DOCUMENT RESUME

ED 356 669 FL 021 179

AUTHOR Stolz, Walter; Bruck, Margaret
TITLE A Project To Develop a Measure of English Language Proficiency. Final Report.
INSTITUTION Center for Applied Linguistics, Washington, D.C.
SPONS AGENCY National Center for Education Statistics (DHEW), Washington, D.C.
PUB DATE 15 Jun 76
CONTRACT 300-75-0253
NOTE 290p.; For a related document, see FL 021 178. Appendixes 9-20 not attached.
PUB TYPE Reports - Descriptive (141)

EDRS PRICE MF01/PC12 Plus Postage.
DESCRIPTORS Adults; Age Differences; Bilingualism; Children; Data Collection; English (Second Language); *Language Proficiency; *Language Tests; *Limited English Speaking; *Measurement Techniques; National Surveys; Program Descriptions; *Questioning Techniques; *Research Methodology; Test Construction; Test Validity

ABSTRACT

A project to develop a measure of English language proficiency (MELP) for use in a national survey of income and education, to estimate the number of people of limited English proficiency, is reported. The preferred form of the instrument was a short series of questions to be asked by an interviewer and answered by an adult member of a household. Principle activities in the project were: (1) development of possible MELP questions; and (2) criterion instruments against which to validate them; (3) field-testing of MELP questions in several ethnic groups; (4) analysis of resulting data to select the best questions for survey use; (5) derivation of scoring keys to translate any pattern of responses into language proficiency categorization; and (6) examination of methodological questions concerning surveys of populations whose native language is not English. A set of about 10 questions were selected for inclusion in the MELP, based on high correlation with respondents' performances on a developed test of English proficiency and on school language proficiency classifications. Slightly different questions were chosen for adults and children. The report describes, in some detail, the procedures used, results obtained, and conclusions drawn. Appended materials include data from analyses, and other supplementary material. (MSE)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED356669

A Project to Develop a Measure of English Language Proficiency

FINAL REPORT

to the

National Center for Education Statistics under Contract #300-75-0253

Submitted by

Walter Stolz and Margaret Bruck

June 15, 1976

Center for Applied Linguistics
1611 North Kent Street
Arlington, Virginia 22209

FL 021 149

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☒ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

BEST COPY AVAILABLE

Abstract

This project was to develop a Measure of English Language Proficiency for use in the Survey of Income and Education (SIE), a large scale national survey to estimate the number of people who are of Limited English-Speaking Ability (LESA) as defined in P.L. 93-380. The preferred form of the instrument was a short series of questions to be asked by an interviewer and answered by a single adult member of a household. Principal activities in this contract were (1) the development of possible MELP questions and (2) criterion instruments against which to validate them. (3) Field-testing the MELP questions in various ethnic groups. (4) Analysis of the resulting data to select the "best" MELP questions for use in the survey. (5) Derivation of "scoring keys" by which to translate any pattern of responses to the MELP questions into a categorization of either LESA or non-LESA. (6) Examination of two methodological questions relative to surveying populations whose native language is not English.

(a) Can a single household respondent give accurate data about all other members of his household?

(b) What differences exist between data collected by monolingual English-speaking interviewers and those collected by bilingual interviewers who are members of the same ethnic group as the respondent?

A set of approximately ten questions were chosen for inclusion in the MELP on the basis of their high correlations with respondents' performances on the developed test of English proficiency and their school classifications as being either LESA or not. Slightly different sets of questions were chosen for adults and children. Discriminant functions were derived using the responses to these MELP questions as discriminant functions yielded a classification accuracy of 75% - 80% when matched against the criteria in a population which had 58% LESA individuals.

An alternative approach to scoring the MELP questions consisted of simply defining certain response patterns as LESA and all others as non-LESA. Such an approach yielded accuracies similar to those of the discriminant functions.

It was found that responses given by a household respondent about others in his household were generally in agreement with those given by the individual himself -- except for a slightly higher incidence of "don't know" responses on the part of the household respondent. Data collected by monolingual English interviewers were generally found to be indistinguishable from data collected by bilingual interviewers.

A problem was discovered in the generalizability of any scoring formula derived in the field test to data collected in the SIE because of sampling differences between the two studies. Thus, it was recommended that the scoring formulae be recalibrated using a sub-sample of the SIE sample.

TABLE OF CONTENTS

		Page
Acknowledgments		i
I.	Introduction	
1.	Background	I - 1
2.	Design Specifications for the MELP Instrument	I - 2
3.	Design Specifications for the Research and Development Effort	I - 7
4.	Investigating the Accuracy of Data Given by the Household Respondent	I - 14
5.	The Language Ability of the Interviewer	I - 15
6.	The Organization of this Report	I - 16
II.	Alternative Approaches to Language Assessment	
1.	Background	II - 1
2.	Criteria for Evaluating Tests	II - 2
3.	A Typology for Classifying Tests	II - 3
4.	Language Assessment Instruments in the MELP Project	II - 16
III.	Instrument Development and Refinement	
1.	Development of Discrete Point Tests	III - 2
2.	Development of the Direct Observation Rating Procedure (DORP)	III - 11
3.	The Monitoring System	III - 17
4.	Development of the MELP Questions	III - 23
5.	The Language Group Representatives	III - 26
6.	Role in Instrument Development and Use	III - 28
IV.	Field Testing the Instruments	
1.	The Basic Design	IV - 1
2.	The Accuracy of First-hand Data and "Proxy" Data	IV - 7
3.	The Language Ability of the Interviewer	IV - 7
4.	The Interviewing Procedures	IV - 8
5.	Monitoring of Interviews	IV - 11
6.	Visits to the Sites by CAL Central Staff	IV - 14
V.	Preliminary Analyses: Selection of the MELP Questions	
1.	"Cleaning" the Data Files	V - 1
2.	Relationships of Individual Questions to the Criteria	V - 2
3.	An Evaluation of the MELP Questions: Reports from the Monitors	V - 14

V. Continued

4. Modifications to the "How Well" Questions..... V - 16
5. MELP Questions as Recommended to NCES..... V - 22
6. Definitions of the MELP Variables..... V - 26

VI. The Criterion Variables

1. The Test..... VI - 1
2. The School Lists..... VI - 12
3. The Direct Observation Rating Procedure (DORP).... VI - 18
4. Relationships Among Criterion Measures..... VI - 19
5. The Correspondence Between List and Dichotomized FCTR..... VI - 24

VII. Derivation of LESA Categorization Procedures for Children

1. The Evaluation of MELP-Based Definitions of LESA and non-LESA..... VII - 1
2. Discriminant Analyses: Child Data..... VII - 3
3. Contingency Table Analysis and the Derivation of Explicit Operational Definitions of LESA and non-LESA..... VII - 15
4. Scoring Keys to be Recommended for Use with SIE Data..... VII - 28

VIII. Derivation of Scoring Keys - Adults

1. Description of Adult Samples..... VIII - 1
2. The Analysis Plan for the Adult Data..... VIII - 3
3. Discriminant Analysis: Adult Data..... VIII - 3
4. Derivation of a Scoring Key Through Contingency Table Analysis..... VIII - 12
5. Recommended Scoring Keys for Categorizing Adults as LESA and non-LESA..... VIII - 22

IX. Finding an Unbiased Estimator of the Proportion of LESAs in the U. S.

1. The List Samples..... IX - 1
2. The Distribution of LESAs and non-LESAs..... IX - 3
3. Adjusting the Face-Valid Definitions of LESA and non-LESA..... IX - 7
4. Summary of Recalibration Recommendations..... IX - 10
5. Adults..... IX - 11

X. Accuracy of MELP Data as Reported by a Household Respondent about Another Adult in the Household

XI. A Comparison of Monolingual and Bilingual Interviewers

1. Production Data..... XI - 2
2. Comparisons of MELP and Test Data as Gathered by Monolinguals and Bilinguals..... XI - 4
3. Performance of the Monolingual and Bilingual Data in a Discriminant Function..... XI - 9
4. Summary..... XI - 10

Bibliography

Appendices:

Appendix 1 - Letter to Center for Applied Linguistics requesting a proposal for research and development activities leading to a Measure of English Language Proficiency.

Appendix 2 - Design specifications for MELP by Dr. Burton Fisher

Appendix 3 - Narrative of principal activities of MELP project: June 1975 - June 1976.

Appendix 4 - Tests considered but not used as criterion measures

Appendix 5 - Window Rock analyses

Appendix 6 - Regression Analyses - Children

Appendix 7 - Regression Analyses - Adults

Appendix 8 - Staff Utilization and Technical Consultants

(Appendices 9 through 20 are not attached to the main report.)

Appendix 9 - Final Criterion Test Battery

Appendix 10 - Criterion Tests - Developing Versions

Appendix 11 - Tests not included in the final criterion package

Appendix 12 - DORP - Rating Scale Description

Appendix 13 - A manual for training monitors to code bilingual study questionnaire

Appendix 14 - Early versions of the CQ Questionnaire

Appendix 15 - Reports from LGR Meeting 1, 2, and 3.

Appendix 16 - Household Information Form

Appendix 17 - Summary of the April 6th and 7th MELP Conference

Appendix 18 - Design Data Collection and Analysis of a Field Test of Instruments and Procedures to Measure English Language Proficiency

Appendix 19 - Report on the Project to Translate into Spanish of the Survey of Income and Education

Appendix 20 - Project Staff's Vitae; Consultants' Vitae

Acknowledgements

The MELP project received so much help from so many individuals and organizations that this section could be the longest one of the report if it were to give credit everywhere that credit is due. But let us acknowledge some of the most important sources of assistance. First, of course, we must thank the Bilingual Studies Group of the National Center for Education Statistics. Les Silverman, Dorothy Waggoner, and Vicki Kojcsich gave more than liberally of their time and energies to the project. They were always there when we needed them and they provided crucial focus, guidance, and perspective throughout the project. Second, we express our appreciation to the Bureau of the Census, particularly Earl Gerson and George Grey for their cooperation and willingness to bend normal procedures and time schedules almost beyond recognition in the service of a better product. Third, we thank the staff of our sub contractor, The Research Triangle Institute, particularly Daniel Horvitz, Tyler Hartwell, and Paul Moore. Their patience with our naivete in the ways of collecting and analyzing survey data were much appreciated.

Throughout the data collection phases of this project, we received enthusiastic cooperation first from several hundred families in San Francisco who tolerated our early, often fumbling attempts to measure their English proficiencies and second from the state education agancies of Florida, Texas, Arizona, and California and from the administrations of the Dade County, El Paso, Ganado (Arizona), Window Rock (Arizona), and San Francisco Independent School Districts for providing us with lists of children and adults to sample from.

Beyond the above-mentioned contributions to the project, the individuals to be thanked number in the hundreds. Many of them are named in the lists of Language Group Representatives, technical consultants, and MELP staff in the Appendix. All of them were superb in giving of their time and talents at a moment's notice. The MELP staff, particularly, was called on again and again to work within

impossible time constraints. This led to all-night and all-weekend work sessions in Arlington, in San Francisco, at RTI, and at the various field test sites. We cannot begin to account for the enormous quantities of perseverance, energy, dedication, talent, and loyalty that we received from these individuals; we can only savor the memories of our shared adventures, express heartfelt gratitude, and list each staff member's name here: Jeanne Freeman, Ted Jones, Greg Strick, Minerva Mendoza-Friedman, Amador Bustos, Eddie Fuentes, Anna Lai, Al Rey, Pedro Ruiz, Mike SamVargas, Terry Webb, Ben Zambalas, Ophelia Balderrama, Rick Chambers, Carry Dunnigan, Guillermo Hernandez, Evangeline Kamitsuka, Carolyn Karelitz, Cindy Lindsey, Gloria Lozano, Claire McKenzie, Roberta Mailman, Annie Panlibuton, Peggy Robbins, Bill Sinclair, Jennie Yee, Gil Garcia, Bill Leap, Leann Parker, Alicia Bustos, and Marina Vargas.

Special thanks to to Donna Ilyin for her help with the adult test, to Jack Upshur for being available on a number of occasions to consult, direct, and give therapy, to Hall Yee for his many insights both into research design and into the concerns of the ethnic groups with which we dealt, to Jack Carroll for his statistical advice, and, most of all to Burt Fisher for his guidance, prodding, perspective, and plain-spoken wisdom during all phases of the project.

The production of this report was greatly facilitated by the careful reviews given by Vicki Kojscich, Les Lilverman, Rudy Troike and Burt Fisher. Finally, Shirley Oravitz was a paragon of skill, patience, and cheer in typing, revising, and editing the manuscript.

For those of us most centrally concerned with the MELP, it was more than a project, it was an adventure, and we thank everyone involved for making it a positive one.

I. Introduction

1. Background

Section 731(c) of Title VII, the Bilingual Education Act, Section 105(a) of P.L. 93-380, the Educational Amendments of 1974, mandates a report on the condition of bilingual education in the nation, including:

> (1) "A national assessment of the educational needs of children and other persons with limited English-speaking ability and of the extent to which such needs are being met from Federal, State and local efforts, including (A) not later than July 1, 1977, the results of a survey of the number of such children and persons in the States, and (B) a plan, including cost estimates,.....for extending programs of bilingual education and bilingual vocational and adult education programs to all such preschool and elementary school children and other persons of limited English-speaking ability, including a phased plan for the training of the necessary teachers and other educational personnel necessary for such purposes;....and
>
> (4) "An assessment of the number of teachers and other educational personnel needed to carry out programs of bilingual education under this title and those carried out under other programs for persons of limited English-speaking ability...."

The survey mentioned above was assigned to the National Center for Educational Statistics and the decision was made to implement it in conjunction with another mandated survey, this one of the number of school aged children in poverty mandated in Section 822(A) of P.L. 93-380. This latter survey was assigned to the Secretary of Commerce (Bureau of the Census) and the "bilingual" survey was "piggy-backed"

onto it. Concretely, this meant that both economic and language questions would be asked of a single very large sample of households. A basic sample of about 155,000 households was designed so as to yield adequate accuracy for the economic data; and an additional sample of 35,000 households was chosen to supplement the main sample to assure a reasonable accuracy level for the English-speaking ability information in each state. This yielded a total sample of 190,000 households to be screened for language data. Finally, a number of questions about health and welfare programs were added to the questionnaire by the Office of the Secretary of HEW. The entire survey effort was named the "Survey of Income and Education" (SIE) and was scheduled to be conducted in Spring, 1976. In order to meet their own production schedule, Census set a deadline of October 3, 1975 for NCES to submit to them the bilingual section of the SIE instrument.

In May, 1975, CAL received a letter from NCES requesting a proposal for research and development activities leading to such a measure of English language proficiency (MELP). Accompanying the letter was a set of design specifications for the project which had been submitted to NCES on March 24, 1975 by Burton R. Fisher, Professor of Sociology of the University of Wisconsin. CAL's proposal was to be submitted to NCES no later than May 15. Both the letter and Fisher's design specifications are appended to this report. (Appendix 1 & 2)

2. Design Specifications for the MELP Instrument

The MELP to be developed had to satisfy two broad criteria: first, it had to be an acceptable and valid measure of English proficiency as that construct is defined in the relevant legislation, and second, it had to be usable within the context of the SIE, a large-scale personal interview survey conducted in households. Each of these criteria will be elaborated and their implications discussed below.

The Construct of Limited English-Speaking Ability

The objective of the survey was to enumerate, in each state, persons who were to be considered of "Limited English-Speaking Ability" (LESA). Section 703 of P.L. 93-380 provides a definition of LESA as follows:

> "Sec. 703. (a) The following definitions shall apply to the terms used in this title:
>
> "(1) The term 'limited English-speaking ability', when used with reference to an individual, means --
>
> "(A) individuals who were not born in the United States or whose native language is a language other than English, and
>
> "(B) individuals who come from environments where a language other than English is dominant, as further defined by the Commissioner by regulations;
>
> and, by reason thereof, have difficulty speaking and understanding instruction in the English language.
>
> "(2) The term 'native language', when used with reference to an individual of limited English-speaking ability, means the language normally used by such individuals, or in the case of a child, the language normally used by the parents of the child."

Fisher further defines the construct as follows:

> The phrase "...<u>speaking</u> and <u>understanding</u> instruction in the English language..." is interpretated to mean <u>oral production</u> (encoding in speech) and <u>aural comprehension</u> (decoding others' speech) in English. In the several education statutes, when reading and writing have been in mind the sophisticated statute drafters have seen fit to specify them directly; such specification is absent here. (Fisher, Pg. 3)

The MELP to be developed for use in the survey needed to relate as directly as possible to the legislatively-defined LESA construct. Thus, the MELP was to have the following characteristics:

1. MELP was to measure English proficiency only: not proficiency in any other language nor language dominance.
2. It did not need to measure reading and writing skills -- nor could it assume them to be present.
3. It had to be targeted on speaking and comprehension skills as required in educational settings.

The Population Relevant to the MELP. The legislative definition quoted above, when viewed from the perspective of a survey, implies a two-stage determination of limited English speakers. The first is to isolate the pool of potential LESA individuals as defined in the Bilingual Education Act. These are persons who were not born in the U.S. or whose native language is not English or who come from an environment where a language other than English is dominant. Satisfying at least one of the above conditions is necessary but not sufficient for a person to be classified as LESA. The second stage is to determine in the survey which of the potential LESAs actually would "have difficulty speaking and understanding instruction in the English language" because of their non-English background. Thus, the SIE was pictured as containing a set of "screening items" which would determine whether a person qualified as a potential LESA individual (i.e. had a background involving a non-English language). If so, then the MELP was to be obtained for that person, and if not, the MELP part of the SIE would be skipped for that person. Fisher says of the screening questions:

> The formulation of these "screening" questions is not a simple matter at all, and there is considerable controversy as to the nature of language questions in Census work. (See Lieberson, 1966, and others.) Under these circumstances, it would be highly desirable that this set of questions be

prepared by the R & D contractor in close association with Census people. (p. 2)

As a pre-test of the screening questions, NCES added a "Survey of Languages" to the July, 1975 Current Population Survey -- a monthly national survey of about 45,000 households taken by the Census Bureau for the Bureau of Labor Statistics. Those questions concentrated on probing for languages other than English present in the household and the native language backgrounds and ethnic origins of the household members. Thus, our project's primary responsibility was to develop the instrument to be used in the second stage of LESA identification; however, the first stage screening questions were also clearly a matter of importance to us.

With respect to the range of ages that the instrument must cover, Fisher concludes:

> Other references in P.L. 93-380 (to preschool education; to auxiliary and supplementary programs for parents of LESA pupils; to elementary and secondary education; to bilingual education under the Adult, Vocational and Higher Education Acts), and the language of Sec. 731 (c) mandating this survey make it clear that the "individual" referred to above may be of any age. However, individuals aged 5 - 17 seem to be of special interest. (p.2)

Constraints as to the Form of the MELP

Fisher was quite specific in characterizing the constraints that the necessities of the Census Bureau imposed on the form of the MELP:

> Census people say that if measurement of LESA is to be carried out in the Census survey, at least four constraints must be observed.
>
> a. "Testing" in any overt form, identifiable by respondents as such, is definitely excluded; this applies especially to "paper-and-pencil" tests.

This places a limit on the kinds of response-eliciting stimuli which can be used to get at LESA.

b. Also categorically excluded is electronic recording of what the respondent says, for later analysis and coding. This places a limit on the kinds of responses to be recorded and the locus of assessment of these responses.

c. A third explicit constraint: LESA measurement procedures must not break rapport during the interview, must fit "naturally" into the context and content of a CPS-like interview (face-to-face or via telephone), and must be within the capacity of its usual CPS and CPS-like interviewers. (On the whole, the latter are women 35 - 40 years of age, with a high school education.) The procedures must not disrupt them.

d. The strong preference of the Census staff is for as simple a measure as is feasible, with a small series of direct questions, answerable by the usual respondent for the household about all of the other members of the household. (In about 60% of CPS interviews, this is the mother.) That is, the preference is for enumeration of the household members, without sampling within the household to select the actual respondents.

This is a strong Census preference, not an absolute requirement. Whether this preference can be gratified, given the need for an adequate measure of LESA (a key NCES requirement), is an empirical question to be answered in the course of R & D work. (p. 1)

Acceptability of the MELP. NCES recognized that if the results of the survey were to be useful to the Congress, they must have the support of a number of concerned constituencies; thus the measure itself must also be accepted as viable

by those constituencies. They included at least: the various non-English speaking minority group organizations, the educational community, and the research community. Therefore, a vital requirement of the project from the beginning was to obtain meaningful input and response from all interested parties at all stages of the work.

3. Design Specifications for the Research and Development Effort.

In broadest outline, the project had two objectives. One was to pick the best MELP possible from among the alternatives which conformed to the specifications outlined above, and the other was to gather validity information to indicate the instrument's strengths and weaknesses. Given the very brief time schedule, it was clear from the beginning that both objectives had to be pursued more or less simultaneously.

Alternative Forms of the MELP - In Chapter II of this report the various approaches to language proficiency assessment will be considered in detail, but it is appropriate here to at least outline the range of techniques available.

Fisher discusses several kinds of MELPs that might fit Census' specifications.* One is simply to ask the Household Respondent about the English proficiency of each individual in the household in a very direct way. Such questions might involve direct ratings of proficiency as well as information about the situations in which each person normally uses English and his history of contact with the language. What literature does exist on this topic indicates that the answers to such questions may be highly correlated with more conventional measures (tests) of English proficiency (cf. Scott, 1973; Bowen, 1974; Capco and Tucker, 1970; and Fishman, Cooper, and Ma, 1971).

A second approach discussed by Fisher that the interviewer assesses the individual's proficiency on the basis of his behavior in the interview. Given the ban by

* In this report the term MELP will be reserved for indicating an instrument for identifying LESA individuals within the context of the SIE.

Census on testing or tape recording in the interview situation, this boils down to the interviewer making a rating of the respondent's English proficiency as displayed in the course of the interview or scoring the presence or absence of specific linguistic features in the respondent's speech. Fisher puts it this way:

> If direct questions about how well an individual speaks and how well an individual understands English, put to that individual or to someone else about him, yield unsatisfactory MELP data, there is an alternative approach. The individual's speaking and understanding behaviors may be observed during the course of the interview itself, in response to questions which at least <u>overtly</u> do not appear to attempt to elicit either a range of language behaviors or an assessment of language behavior by the respondent . . . The interviewer may be trained to record and assess/rate behaviors he has been cued to watch for on forms developed by the R & D work on MELP. This is <u>not</u> unusual procedure in good psychological and social research and in assessment work in organizations. People without previous expertise and special qualifications have been successfully trained to make reliable and accurate reports and assessments of behaviors during group interactions and individual performances, in field and in laboratory situations. (p. 3)

A serious implication of this approach would be that the interviewer would have to talk with each person who was rated. This would undoubtedly call for some sort of within-household sampling and a significant reduction in the total number of individuals for which LESA and non-LESA categorizations could be obtained because of the greater cost of directly interviewing more than one respondent within a household.

Criterion Instruments - But the needs of this project extended beyond instruments which could possible qualify as SIE MELPs since a primary purpose of the R & D effort was to validate such a MELP, and that implies validating it against some other instrument -- presumably a more direct, accurate, or widely accepted measure -- which could serve as a criterion during field testing. While such instruments did not have as an absolute requirement the restrictions on form imposed by the Census Bureau, (since they were to be used only in our field test) there were severe logistical constraints on what could be used because of the scope and time schedule of the field test activities. In particular, since the objective of any field test would be to try out an instrument under conditions similar to those of its eventual use, the field test had to be household-based, and thus the criterion measure(s) had to be usable in a household setting. This would seem to eliminate assessment procedures involving costly and/or delicate equipment. Also, the criterion measures had to be applicable to people of all ages and from all ethnic-linguistic groups. None of the measures could assume reading or writing skills on the part of the respondent. Given all of these constraints, criterion instruments had to measure as directly as possible language functions necessary for success in educational settings.

Validation - Fisher offers the following discussion of validation vis à vis educational criteria:

> (a) On validity; MELP is to measure what it is intended to measure -- the characteristics and relative proficiency of "speaking and understanding instruction in the English language," which make a difference or could make a difference in the individual's progress in a course of education or training. How "limited" ESA is, for present purposes, is to be referred against the language performance of individuals whose ESAs are seen by the schools as barriers of varying strength to effective learning, when instruction is in English.

(b) This applies to individuals in (preschool?), elementary, secondary, postsecondary, adult and vocational education programs. MELP validity studies ideally should be carried out in all of these contexts.

(c) It should be recognized that different educational agencies (SEAs, LEAs), schools and programs use different measures and criteria (of different worth in terms of scientific standards) both of ESA and of effective educational progress. The procedures for identifying individuals for whom LESA is in varying degrees a barrier to utilizing effectively instruction in English will thus also differ. The R & D contractor may be able to make some choices among these educational sites, as to where MELP developmental and validation studies should be carried out. (The modes of stratification for a purposive sample of sites in which to carry out such studies is left for later consideration by the R & D contractor.)

(d) For both practical and theoretical reasons, we are not likely to arrive at a "true" (essentially metaphysical) definition and measure of characteristics and degrees of ESA which universally ought to facilitate or inhibit educational attainment. We can obtain administrative identifications, in the schools as they are and by the identification methods they currently use, of individuals inhibited from normal educational attainment by LESA. This is a ubiquitous problem in research on exceptionalities, and the approach suggested here echoes experiences derived from that research. (p. 4)

Fisher's suggestion of using schools' administrative identifications of LESA and non-LESA individuals is an important one for two reasons: First, given that the present state of the art of language assessment is in its infancy (see Chapter II)

amd thus elicits little agreement from specialists about what is the single "best" approach, then a logical reference is to those actually making assessments in a routine way, however it is being done. If it is unclear what is the best approach, then a viable objective is to simulate the most enlightened among the currently used practices. Second, since administrative identifications are typically used for making decisions among a small number of alternatives, they are usually procedures yielding discrete and often binary classifications. On the contrary, most non-administrative assessment instruments yield scores that are basically continuous in nature and do not lend themselves to making dichotomous classifications without considerable arbitrariness. Thus, school's administrative screening procedures for non-English speaking students were to play an important part in this project.

The general strategy employed with respect to validity was to focus on content validity and on concurrent validity. Content validity was addressed first by recruiting a staff with expertise in test development and linguistics and who also were drawn from a number of ethnic-linguistic groups. Second, we asked a number of specialists in the areas of language and language testing who were not otherwise associated with the project to comment on the adequacy of both criterion measures and possible MELPs. Third, CAL convened a large board of "Language Group Representatives" to criticize early versions of all instruments and to make suggestions about how they could be improved to be more "culture-fair" relative to each representative's group.

Concurrent validity was obtained by eliciting data from field test respondents on <u>several</u> "criterion" measures of English proficiency, each representing a particular approach to language assessment. (As it will be seen, at least as much effort initially went into the development of appropriate criterion instruments as went into the development of the MELP itself.)

The distinction between a valid measure and an accurate estimator. Although finding a valid MELP was an important objective of the project, the overriding objective of the MELP itself is important to keep in mind; that is, to accurately estimate the proportion of LESA individuals in the country. There is a crucial but subtle difference between validating a measure of English language proficiency and constructing a procedure to estimate the proportion of limited English speakers in the country. When validating a new measure, one correlates it with "criterion" measures of the same construct measures which are already established or are more direct measures than the one to be validated. The important issue in validation is the extent to which the candidate measure tends to agree with (give the same answer as) the criterion measure(s) on a person by person basis across a large number of respondents.

On the other hand, when constructing an estimator of a population parameter, it is most important that the estimator performs accurately at the level of the population. Thus, if the "true" proportion of LESA individuals in a given population is 0.2, the crucial property of a successful procedure for estimating that quantity, is that it gives a value of about 0.2. Whether or not the estimator classifies the "correct" 20 percent of the population as being LESA is a secondary consideration. For example, consider the following three tables involving mythical populations of 100 persons each.

Table 1:

Candidate estimator	"true categorization" LESA	-LESA	Total
LESA	20	0	20
-LESA	0	80	80
Total	20	80	100

Table 2:

Candidate estimator	"true categorization" LESA	-LESA	Total
LESA	0	20	20
-LESA	20	60	80
Total	20	80	100

Table 3:

Candidate estimator	"true categorization" LESA	-LESA	Total
LESA	20	20	40
-LESA	0	60	60
Total	20	80	100

In Table 1, we have the best of all situations in that there is perfect agreement between the candidate estimator's categorization of the 100 individuals and their respective "true" categorizations. This estimator is thus a perfect estimator (it estimates the same percent of individuals to be LESA as is the true case), and it is also perfectly valid (every individual is assigned to the correct category). In Table 2, however, the estimator is an accurate estimator, since it gives the correct proportion of LESA persons in the population, but it is not particularly valid in the sense that it gives the correct categorization for only 60 of the people. Finally, in Table 3, the estimator is relatively valid -- giving the correct classification for 80 of 100 people -- but a poor estimator since it over estimates the number of LESAs in the population by 100%. While very high validity in the above sense is desirable, because it implies an accurate estimator, we must never forget what our ultimate objective is: to produce a good estimator of the proportion of LESA individuals in the nation. It is conceivable, then, that this project could find a MELP which is not highly valid as compared with available criterion measures -- all fallible to be sure -- yet which is a reasonably accurate estimator in the sense of closely matching the proportion of LESA individuals in a population as given by one or more of these criterion measures. The situation is

somewhat curious if the true proportion of LESAs in a population is quite small, say 10%. Then, a MELP which simply declared everyone to be non-LESA would be 90% valid; however, it would be nonsense as an estimator of the proportion of LESAs. On the other hand, a MELP operating in such a population might display a validity of 90% or less while estimating the true proportion of LESAs quite closely. It would achieve this by falsely categorizing approximately equal numbers of LESA and non-LESA individuals. Generally speaking, we will evaluate all MELPs both in terms of validity and accuracy of estimation. The former will be indexed simply by the proportion of a sample categorized the same by both MELP and the criterion measure against which it is being compared. The latter will be indexed by a quantity to be called "% bias" (see Chapter VII), which will be a function of the difference between the proportion of the sample identified as LESA by the MELP and that identified as LESA by the criterion.

4. Investigating the Accuracy of Data Given by the Household Respondent.

An important requirement of any MELP questions which were to fit Census' desired guidelines was that one adult in the household (the Household Respondent) had to provide accurate answers to the questions for every member of the household. This matter was investigated within our study in the following way: The interviewer was told to follow "standard" Census Bureau interviewing procedures in the sense of beginning each household interview by locating a responsible adult who was willing and able to act as the Household Respondent and to provide information about another member of the household. While in the SIE questions would be asked about all others in the household, in the present study our focus was on only one designated individual -- generally a child or adult whose name we had received from the local school. The procedure was then to ask all Census-type questions of the Household Respondent about this Designated Respondent. Then, if the Designated Respondent was an adult,

the questions were also asked of him directly about himself. Although our interviewers collected some questionnaire data directly from child Designated Respondents, they were not analyzed because Census did not plan to collect such information from children under any circumstances. Therefore, all questionnaire data collected on children in this study can be considered to have been provided by an adult Household Respondent and thus qualifies as essentially "proxy" data. On the other hand, every adult Designated Respondent in the study provided questionnaire data about himself, and this "first hand" data formed the basis of all analyses of adult MELP data reported in Chapters V and VIII. In addition, proxy data were collected from a Household Respondent different from the adult Designated Respondent when such an individual was available at the time of the interview. In single-adult households, the adult Designated Respondent and the Household Respondent had to be the same person and thus proxy data were simply not available for that individual. The relationship of the proxy and first-hand data for adults is discussed in Chapter X.

Of course, all criterion instruments were administered directly to the Designated Respondent.

5. The Language Ability of the Interviewer

Another concern about the accuracy of the data revolved around the fact that monolingual (English speaking) interviewers would inevitably be dealing with respondents whose English proficiency ranged from excellent to none. And, in addition to the linguistic factor, there was also the cultural difference between the monolingual, probably Anglo, interviewer and the ethnically distinct respondent. This difference could easily take its toll in refusals to be interviewed or on the rapport between the two and thus influence the character of the data collected. In order to evaluate the severity of these problems, one component of the design of the field test was to compare the data collected by monolingual (English) interviewers

and bilingual interviewers whose native language and ethnic origin was that of the respondent's. This was done by matching the assignments of monolingual with bilingual interviewers in each site through randomizing the names and addresses of the individuals they were to interview. Monolingual interviewers were given standard Census instructions, that is, if communication with the Household Respondent was severly impeded by the respondent's lack of English proficiency, the interviewer was to find someone else in the household or neighborhood who could act as a translator. Bilingual interviewers were instructed to conduct their interviews in English whenever possible and to refer to the native language only when absolutely necessary. They were encouraged to consult informally with one another in advance about the proper translations of various questions, but no formal, written translations of the questions were used.

6. The Organization of this Report

In subsequent parts of this report, the project's activities will be described in the following order:

1. A review of the various approaches to measuring language proficiency (Chapter II).
2. The instrument development activities -- both of possible MELPs and various criterion measures (Chapter III).
3. The field test in which the instruments were used in several ethnic-linguistic communities (Chapter IV).
4. The selection of the MELP questions for recommendation to NCES (Chapter V).
5. Analyses of the criterion measure data, particularly focusing on the relationships among the measures (Chapter VI).
6. Construction of scoring keys for children and adults by which individuals could be categorized as LESA or not on the basis of their responses to the

MELP questions (Chapters VII and VIII).

7. Observations on generalizing the results of Chapters VII and VIII to determining LESA and non-LESA categorizations for individuals surveyed in the SIE (Chapter IX).

8. Investigation of the validity of the MELP data provided by a Household Informant about other adult members of the household (Chapter X).

9. Investigation of "interviewer effects", comparing the data collected by interviewers who are from the same ethnic-linguistic community as the respondent with data collected by monolingual English "Anglo" speakers. (Chapter XI).

II. Alternative Approaches to Language Assessment

1. Background

In the past decade, the nature of language assessment has changed as a result of a shift of emphasis in current linguistic theory from structuralism to functionalism. Through the 1960's, the language tests reflected the view-point of the structural linguists (cf. Chomsky, 1965): that language is grammar-based and can be divided into such subcomponents as phonology, syntax, and semantics. English language proficiency tests were constructed to measure the individual's knowledge of a number of these structures.

In the late 1960's, some linguists (e.g. Hymes, 1967, Labov, 1970) emphasized that knowing a language involved more than being able to conform to its rules of syntax, phonology, and vocabulary; it also included being able to use language in communication situations. The speaker had to demonstrate that he knew when to speak, to whom he should speak, where he should speak and how he should speak. Functional and communicative aspects of language were stressed. The individual's ability to appropriately express himself and make himself understood were examined. Test constructors emphasized the importance of collecting data in "natural" or contextually relevant situations. Instruments were developed to assess global communication skills in specific types of contexts (e.g. the classroom) rather than a number of specific grammatical, phonological, and semantic skills in a generalized or unspecified context.

This drift in both theoretical and measurement emphases illustrates how tentative the linguist's hypotheses are about the nature of language. It is most important to recognize this tentativeness when evaluating the adequacy of language proficiency tests, since different test developers may have rather different conceptualizations about the nature of the phenomenon that they are attempting to

measure. For example, the Illinois Test of Psycholinguistic Abilities (Kirk, McCarthy, and Kirk, 1968) was created within the context of Charles Osgood's theory of language. Within any other frame of reference -- e. g. most linguists' -- the test is of dubious validity. In the present case, a test may be reasonably valid to a person viewing language proficiency as tacit knowledge of an isolated set of syntactic, phonological, and semantic rules, but it may be quite beside the point for someone viewing language as the ability to perform "appropriately" in a set of communication situations. Only after agreement is reached on what is to be measured can one set about evaluating the effectiveness of various measurement approaches. In terms of "validity", as the term is used by psychometricians, we have a situation where "experts" may not agree on the construct validity of a given instrument because they do not agree on the construct itself. Such a condition essentially precludes the existence of any universally accepted measure or test of the construct, and this is exactly what a review of the language testing literature shows, i. e., that there is considerable disagreement among specialists about which of the hundreds of existing tests are "the best." Even within the slightly more restricted domain of educational settings there is still little consensus on "the best" instrument.

2. Criteria for Evaluating Tests

Assuming, however, that some agreement can be reached about the nature of the phenomenon to be measured, it is useful to set up some criteria that an "ideal" measure of English proficiency should meet. We propose the following six criteria:

1) The test should be a broad measure of English proficiency in the sense that it should measure productive (speaking) as well as receptive (listening) skills. For older children and adults it should also measure proficiency in reading and writing (a criterion not relevant to the present application).

2) The test should reflect differential proficiencies in different domains of use (e.g. home, school, church, peer, adult, etc.). (Again, for the purpose of this project, the test need only be a measure of proficiency in one domain: the school setting).

3) The test should be reliable and valid, a universal requirement for any test. It should have high construct, content, and face validity.

4) The test should yield scores that are readily interpretable relative to the objectives of the testing. Usually this means that norms must be available for groups similar to those with which the test is to be used. If the test has been constructed as criterion-referenced or performance-based, then norms are not necessary, provided that scores are interpreted as intended by the test constructor. In some applications, where all comparisons and interpretations of scores are done internally to the study (as in the present project), norms are not necessary because comparisons of persons inside the study are not being made with persons outside the study.

5) The test should be easy to administer in a reliable fashion.

6) The test should be easy to score in an unequivocal fashion.

3. A Typology for Classifying Tests.

In this chapter a number of English language proficiency tests will be reviewed and evaluated relative to the six criteria of the previous section. Each of these tests is currently in use with adults and children from non-English or bilingual backgrounds. In order to facilitate this review, however, the tests will be cast into a four-fold typology. As will become clear, tests which are members of the same type tend to share similar strengths and weaknesses relative to the criteria. Thus, a number of important attributes of a test can often be identified simply by placing it in its appropriate category.

The four categories are actually the conjunction of two independent dimensions. These will be explained briefly and then more extensively as the tests themselves are discussed.

The first dimension is labeled discrete-point vs. integrative and refers to the assumptions and intents of the test constructor and the test user. A discrete-point test is one which attempts to analyze English proficiency into its atomic components and then test each of the components separately. This approach was typical of the structural linguists of the 1950's and early 1960's who believed that to test language proficiency one tested knowledge of the facts of the language: e.g., syntactic rules, morphology, vocabulary, etc. The specific format of the test was important only in that it should facilitate revealing the knowledge and not impede it. (For example, the test format should not, in itself, place a heavy load on memory or call on large amounts of non-linguistic -- and thus irrelevant -- knowledge and abilities, e.g., intelligence.) The crucial feature, though, of the discrete-point approach is the assumption that if one is "proficient" in knowing enough of the components of a language, he is proficient in the language. In a sense, a discrete-point test is a collection of mini-tests, each testing a separate sub-construct and fielding a profile and summary measure of language proficiency.

An integrative test is one which involves a task assumed to call upon a large range of the phenomena under examination. The degree to which that task is accomplished becomes the score on the test. For example, taking dictation is considered by many specialists to involve a large range of linguistic skills, both receptive and productive. An integrative test then might be to dictate a passage to a respondent and simply count the number of errors he made in his transcription. An integrative test is assumed to index the respondent's integrated English proficiency rather than the separate components of his proficiency.

The second dimension deals with the relevance of the assessment situation to the behavior of interest, and it is called the direct-indirect dimension. A

direct test or assessment is one which samples directly from the behavior to be evaluated. For example, if one is interested in English proficiency in the classroom, a direct assessment would be to observe the respondent in his routine classroom activities and then in some way rate or score his performance in that situation. As the evaluation situation becomes more contrived and/or different from the situation of interest, the test becomes more indirect. Notice the implicit assumption here is that the evaluation is not of traits or abilities or knowledge residing entirely within the respondent. Rather, the evaluation is of the individual's abilities to interact with his environment in specified classes of situations. This is a thoroughly appropriate assumption to make in the present project given the legislative definition of LESA as being "difficulty in speaking and understanding instruction" because of a non-English language background.

Since directness is a joint property of a test and what it is meant to measure, a test is neither direct nor indirect in and of itself. It may be very direct when used to measure one sort of behavior and indirect when measuring another. Valid direct tests are face-valid and construct-valid while indirect tests must generally depend on the establishment of concurrent validity in order to be considered valid. Also, it is clear that the direct-indirect distinction is in fact a continuum and that tests are not direct or indirect in any absolute sense, but only more or less direct.

Indirect-Discrete Point Tests. These tests can be sub-divided into two groups: standardized and non-standardized.

Two examples of standardized discrete point indirect tests are: Test of English as a Foreign Language or TOEFL (ETS, 1975) and Michigan Test of Language Proficiency (Upshur, et al, 1964). The Michigan test is designed to be a test of English language proficiency for adults enrolled in college and is composed of three

sections: grammar, vocabulary, and reading comprehension. It measures such language facts as: word order, noun and pronoun forms, verb tenses, modals, ellipsis, prepositions, and idioms.

The TOEFL was also designed to measure the English proficiency of foreign students applying for college admission into the U.S. It is composed of several sections: Listening Comprehension, English Structure, Vocabulary, Reading Comprehension and Writing Ability. Items on these subtests are designed to measure specific language facts.

Many of the unstandardized indirect discrete-point tests are pilot tests for which later refinement and standardization are planned. Three are discussed: Bilingual Syntax Measure (Burt, Dulay and Hernandez-Chavez, 1974), the MAT-SEA-CAL (Matluck and Matluck, 1975), and the Ilyin Oral Interview (Ilyin, 1972).

The Bilingual Syntax Measure tests a child's (ages 4 to 9) ability to produce specific grammatical structures in English (or Spanish) which are supposedly important indicators of structural proficiency. The child is shown a picture, and is asked a specific question about it. The question is so phrased as to elicit a specific grammatical structure.

The MAT-SEA-CAL was designed to measure a child's ability to understand and produce distinctive characteristics of English. The three sections: Listening Comprehension, Sentence Repetition, and Structural Response test specific phonological, morphological, syntactic, and lexical items.

The Ilyin Oral Interview is a test of oral English language proficiency for adults (from 13 years on). The examinee is asked to give complete statements in response to a series of questions based on a sequence of pictures. Answers are scored separately for information conveyed and grammatical elements. As in the other two tests the questions are structured so as to elicit specific grammatical structures.

The above three tests have been classified as examples of "indirect tests"; in that while the language testing situation is probably closer to "real-life" than that of the standardized tests previously discussed, they do not represent or directly sample from naturalistic situations. That is, in normal discourse while we might ask people questions about pictures, we do not structure questions to elicit specific linguistic forms, nor do we ask a string of 28 consecutive questions. Thus, these instruments are thoroughly "test-like" and bear little resemblence to normal dyadic interactions, even between students and teachers.

The test constructors of the three example tests described above all state that norms, reliability and validity for these tests are forthcoming.

One additional test (or technique) should be mentioned in this section: imitation tests. Here the task is for the examiner to say a specific sentence (one long enough so that the examinee can't memorize it) which the examinee then is to repeat verbatim. The rationale for this technique is that correct repetitions indicate underlying knowledge of the structure of the sentence. Although there is no single generally accepted imitation test, it is easy enough for a test-constructor to draw up and use a list of sentences which contain the important "language facts." Examples of this approach are Naiman (1974), Menyuk (1963), and Natalicio and Williams (1970).

How well do these types of tests meet the six criteria proposed for an "ideal" language proficiency test? First, the tests vary in terms of the range of language skills they assess. Some (TOEFL) assess reading, writing, and listening comprehension, while others purport to test only oral skills (Ilyin, B.S.M.). However, there does not appear to be one test that measures all four language skills (speaking, understanding, reading and writing). Secondly, it appears that all these tests focus on one variety of language: formal standard English.

Third, while the standardized tests have norms and assessments of reliability and concurrent validity attached to them, they and all indirect discrete-point tests have recently been called into question because of the assumptions underlying them. Critics take issue primarily with the assumption that language proficiency is simply the tacit knowledge of a collection of "facts" about the language which can be tested for, one by one, outside of any context in which the respondent would normally use the language. Clearly, both the concept of discrete-point testing and the indirect nature of most discrete-point tests are under attack. (For a summary of these criticisms, see Jones and Spolsky, 1975; Upshur, 1971.)

The main advantages of indirect discrete point tests are that they are comparatively easy to administer and score.

Direct Discrete-Point Tests. The main differences between this set of tests and those described in the previous section, are in the techniques used to elicit the individual's responses. Because these types of tasks attempt to elicit language in "natural" situations, the responses are usually strings of sentences, rather than single sentences or words. However, the tests are considered discrete-point in that analysis of the subsequent responses involves counting and analyzing specific structures which the test-constructor states are important subcomponents of language proficiency. Two examples of these tests, the Basic Inventory of Natural Language (Herbert, 1975) (BINL) and the Language Cognition Test, (Stemmler, 1975) are tests of productive skills for children. For the BINL, children are trained to talk to each other about pictures. After a number of such training sessions (for which the test constructor must do on-site workshops) the children's subsequent narratives are recorded and analyzed for such features as syntactic complexity, fluency, and sentence length.

The Language Cognition Test is similar to the BINL except that the child talks to an adult about a picture and some familiar objects. The responses are recorded and later analyzed for: basic sentence types, transformations, verb constructions, and adjective types.

The disadvantages of these tests are that they only measure oral production; they have not been validated, or standardized, and there is no information on their reliability. While they may be easy to administer, the scoring procedures are quite lengthy and require some training of the scorer. Positively, these types of tests can be readily used to assess language in many domains. For example, one could construct the elicitation situation in such a way that the subject tells a story to his friend, or to his mother, or to his teacher etc.

Direct-Integrative Tests. The procedure which bests demonstrates a direct integrative assessment of overall language proficiency (oral and written) is the Foreign Service Institute's oral interview and rating technique (FSI, 1963). Here the main emphasis is assessing how well a person can communicate in a language for particular purposes in given situations. Usually the respondent is brought in to converse for a half hour or so with two observers, at least one of whom is a native speaker of the language. The topics and the situations covered generally are chosen to be as similar to typical on-the-job situations as possible. The speaking test ends when the two interviewers are satisfied they have pinpointed the respondent's rating level. This usually occurs within 30 minutes (and frequently within 5 to 10 minutes). The 9 point rating scale ranges from (1) which is defined as elementary proficiency to (5) which is native or bilingual proficiency. Each rating is well defined in terms of the level of language used. For example, the first level (Elementary Proficiency) is accompanied by the following description:

Elementary Proficiency

S-1: Able to satisfy routine travel needs and minimum courtesy requirements. Can ask and answer questions on topics very familiar to him; within the scope of his very limited language experience can understand simple questions and statements, allowing for slowed speech, repetition or paraphrase; speaking vocabulary inadequate to express anything but the most elementary needs; errors in pronunciation and grammar are frequent, but can be understood by a native speaker used to dealing with foreigners attempting to speak his language; while topics which are "very familiar" and elementary needs vary considerably from individual to individual, any person at the S-1 level should be able to order a simple meal, ask for shelter or lodging, ask and give simple directions, make purchases, and tell time.

R-1: Able to read some personal and place names, street signs, office and shop designations, numbers, and isolated words and phrases. Can recognize all the letters in the printed version of an alphabetic system and high-frequency elements of a syllabary or a character system.

Other government agencies have further subdivided the skills and devised rating scales for listening and writing proficiency.

Dealing specifically with the FSI oral interview, how well does it meet the criteria suggested above?

1) The procedure can be used to assess the full range of an individual's oral skills.

2) From the rating descriptions, it appears that many different domains of language use are being assessed (e.g. can order a meal, ask directions). However, it is unclear how well one can assess language use in a variety of domains in such a short time.

3) The inter-rater reliability in the oral interview situation is very high (Clark, 1975). What is not known is whether the measured proficiency of the respondent fluctuates from day to day. Thus he might receive a variety of ratings were he retested on several consecutive days. Also, it should be emphasized that FSI maintains extensive training and recalibration programs for its interviewers. Thus,

this high inter-rator reliability is quite costly.

There are no data on the predictive validity of the test (i.e., how well respondents actually perform "in real life" in a number of sociolinguistic contexts). Constructors of the test state that it is highly face valid; however, many have taken issue with the apparent "naturalness" of the testing situation (e.g., comments in Jones and Spolsky, 1975). It is important to keep in mind that because it is a testing situation (and not a tea-party) it can never be totally natural. Clearly, any time a person knows that his performance is being formally evaluated, the situation becomes somewhat "unnatural" for him.

Lastly while this procedure may be quite easy to administer, scoring tends to be difficult and expensive in terms of interviewer training time and sophistication.

The Dailey Oral Language Facility test (Dailey, 1968) as adapted by Cohen (1975) is an attempt to adapt rating scale procedures for use with children. Here the children are asked to tell stories about different pictures which represent three different social domains (home, school, and neighborhood). The stories are then rated by two raters on a number of 5 point scales (e.g., general ability to communicate, fluency, grammar, pronunciation, rhythm, intonation). This test is similar to the BINL except that the analyses of the data are global. It is similar to the FSI procedure, except that the stimulus situation is more closely controlled.

Generally, oral interview and rating techniques are not widely used outside government agencies for several reasons. The most important reason is that they are very expensive to maintain. As indicated above, FSI interviewers are highly trained specialists who are required to return frequently for retraining and recalibration. Extensive research on language and attitudes has indicated that untrained raters often make highly biased judgements about a person's language ability based on non-linguistic variables (e.g., sex, race, dress, etc.). A secondary reason

is that the use of such a technique in different language use situations (classroom, vocational) and with different age groups would involve completely reformulating the interview procedures and the criteria for evaluating an individual's performance. Thus, the technique is expensive both to maintain and to initiate. In fairness to the approach, it must be admitted that we do not yet know the minimum amount of interviewer training which is necessary to achieve reasonable reliability on various scales used in different contexts. The possibility certainly exists that acceptable results could be obtained in some situations and with some groups using different, less costly training procedures then those used by FSI. Although the Dailey has not been thoroughly developed to date it may be a start in this direction.

Indirect-Integrative Tests. These are tasks which do not have a high degree of face-validity, but purport to measure "global" language proficiency.

One set of tests in this group are termed "reduced redundancy tests" (Spolsky, 1971). The main rationale underlying these tests is that there is a great deal of redundancy in language which is particularly useful to the non-native speaker as he makes guesses about the meanings of utterances that he hears or reads. If this redundancy is removed, it should be much more difficult for him to continue to communicate.

Redundancy can be removed in a number of ways. In the Cloze Test (Taylor, 1953), redundancy is reduced in a reading task by deleting every nth word in a paragraph, and the respondent is required to supply the missing words. Scoring involves counting either the number of exactly supplied words or the number of contextually acceptable responses.

The correlation of this test with other tests of language proficiency is quite high: .83 with the UCLA language proficiency test, .73 with the TOEFL listening comprehension test (Darnell, 1970, Oller, 1972).

Another test of reduced redundancy is the dictation test (with or without noise). In the traditional dictation test (without noise) the person is read the dictation and he writes it down (Gradman and Spolsky, 1975). The number of errors are counted and subtracted from a base line score. Such a test was found to correlate .94 with the UCLA English Language Proficiency test. It also correlates highly with the Cloze test (Oller and Streiff, 1975). It is called a reduced redundancy test in that many of the cues used in natural situations are removed. If a person's internal grammar is incomplete, "the kinds of hypotheses that he will make will deviate substantially from the actual sequences of elements in the dictation." Oller mentions, as a example of this, the student who converted a phrase "Scientists from many nations" into "scientists' imaginations" (Oller and Streiff, 1975).

The reduced redundancy test with noise involves giving the student a number of sentences in the target language which have been masked by the introduction of white noise (Gradman and Spolsky, 1975). The student attempts to write out, or repeat each sentence. This test has been validated against various tests: TOEFL (.75); TOEFL Listening Comprehension (.89), TOEFL Vocabulary (.85) and the Ilyin Oral Interview (.69).

These reduced redundancy tests all share a common set of problems, as well as advantages. The tests are heavily dependent upon orthography (at least in their present forms), and as a result it seems unclear how directly they actually measure oral skills. The tests do not seem well suited for investigating language proficiency in various domains, since it appears difficult to construct these types of tests to measure a person's ability to communicate with a certain person in a specific setting. In most cases the tests seem fairly easy to administer and score. Perhaps the biggest question associated with all integrative - indirect tests concerns their validity. Clark (1975) contends that the ultimate usefulness of such tests will rest on the

magnitudes of correlations between them and more direct measures (specifically FSI type tests). Nevertheless, the evidence provided by concurrent validation with other relatively indirect measures plus their ability to be employed efficiently and economically is encouraging.

Another category of indirect integrative measures includes word naming and word association tasks. Macnamara (1969) defines these as brief economic measures to assess undifferentiated degrees of bilingualism. Because these measures have typically been used to assess degree of bilingualism, they are usually administered in two languages. However, they also can be administered in one language as a test of general proficiency. They have been used to assess language usage in different domains (cf. Fishman, Cooper, and Ma, 1971) and are very easy to administer and score. Their validity will be discussed below.

The last variety of integrative indirect tests to be discussed is that of self-report. Here the subject rates his own language proficiency. Depending upon how the interview questions are structured, he can be asked to rate his proficiency in a number of different domains or situations (church, school, in a restaurant, giving directions). The rating scale itself can be made up of any number of points with as much description or definition of each point as the test constructor cares to make. These scales have the advantage of being very easy to give and very easy to score. There are many unanswered questions about the utility of the rating scale, and the validity of the approach is controversial (see below). It is clear that young children cannot rate their own proficiency, and that parents' or teachers' ratings of children's proficiency might not be valid. For example, teachers' ratings could be influenced by attitudes and stereotypes about the child which are non-language related. We do not know how accurately a parent can rate his child's proficiency in a language if the parent does not see the child use the language and/

or does not know the language himself. Also the ratings might be affected by such variables as humility, and social pressures to respond in certain ways.

As noted above we also have little information on the validity of these rating scales. Arsenian (as cited in Macnamara) cites validity estimates of about r=.80 obtained by correlating a language background questionnaire (a series of questions about respondent's and family's language use in different situations) with ratings of linguistic proficiency made by interviewers. Macnamara attempted to relate a series of indirect measures (language background, self-rating, word naming, reading speed, word detection and word completion) to a number of "more" direct and standardized measures of language proficiency (Gates reading test, a listening comprehension test, a story telling test)*. He used the direct tests as criterion variables and the indirect tests as predictor variables. While he found that the language background questionnaire was not a good predictor of performance on the direct tests, the self-rating scales were powerful predictors. Macnamara had the subjects rate themselves on four different scales (reading, writing, speaking, listening). However, in his analysis, he found that little accuracy was lost by combining the four ratings into one. Of all the indirect measures, he found that self-ratings of "speed of reading" was the most powerful predictor of bilingual skills, this however is probably due to the fact that many of the criterion tests involve this skill. Other indirect tasks contributed in less powerful ways to the prediction of the criterion tests.

In our review of language proficiency tests we realize that we have not provided an exhaustive list of all available measures. Rather we have attempted to sample and furnish a critique of those that are more commonly used and those which show promise of being good measures.

---

* Macnamara was interested in assessment of bilingual proficiency and thus administered the above tests in English and French. He obtained difference scores on each test and correlated these among tests. However, his results are interesting for those concerned with the measure of language proficiency.

4. Language Assessment Instruments in the MELP Project

Possible MELP Instruments. With respect to the MELP (that is, the instrument used to identify LESA individuals in the SIE), the two most tenable approaches have already been mentioned in Chapter I: (1) A set of opinionnaire-type questions to be answered by a Household Respondent about the English proficiency, use patterns, and history of each member of the household, and (2) a direct rating or scoring system completed by the interviewer during the interview. The prohibition by Census of any obvious testing ruled out anything but these approaches. In the second option above, the rating and scoring procedures would have to be designed as essentially covert measurement. That is, the interviewer would assign a proficiency score to the respondent without the respondent being aware that his English was explicitly being assessed. If the interviewer were to simply rate the respondent's proficiency in a way analogous to an FSI rating, it would qualify as integrative and relatively direct. It would be indirect only in the sense that the household interview situation does not obviously sample directly from language use requirements in instructional settings. However, if the interviewer were to observe and code (perhaps on a checklist) a set of features as they occurred during the interview -- e.g. various sentence types, verb tenses, dependent clauses, etc. -- the assessment would qualify as a discrete point direct test. Fisher discusses this approach as follows:

> Specialists in applied linguistics have knowledge of the components and dimensions of phonology (accents, sounds, some dialect features), of lexicon, of syntax and of utterances to be used to characterize oral production and aural comprehension. (Parenthetically: Bilingual interviewers or non-verbal behavioral response indicators may be necessary, where an individual comprehends but does not speak English.) Applied linguists are aware of certain central "diagnostic" linguistic features of adequate and inadequate English language usage and comprehension. If they do not already know which of these linguistic features are most highly correlated with other features of English language usage, they can determine this empirically in R & D work at the educational sites. (The purpose of this is to shorten the list of language behaviors to be observed, for entering into an assessment of ELP made by trained interviewers. The aim is practical -- while maintaining a list of critical items long enough for MELP reliability.) (p.5)

We suspect that Fisher is overly confident of linguists' knowledge of linguistic features that are particularly diagnostic of overall proficiency. It is exactly that "knowledge", as exemplified in discrete point tests, that has recently been called into question by Jones and Spolsky (1975). It is important to note that the rating and the scoring approaches were never thought of as anything but possible last-ditch, fall-back MELPs, to be considered only if ratings by a Household Respondent proved a complete failure. They were considered as such because of their necessitating the interviewer to converse face-to-face with each individual being given a LESA - non-LESA categorization.

Possible Criterion Instruments. With respect to possible criterion instruments -- i.e., instruments to use as standards against which to develop and calibrate the MELP -- the restrictions as to form were somewhat less severe.

Clearly, discrete point indirect tests were prime candidates for the following reasons:

1. They are easy to administer and score.
2. They need not involve paper and pencil.
3. A number of them have been developed, all or parts of which might be usable.
4. While more controversy about their validity is present now than ever before, discrete point tests still have the largest single block of adherants in the testing community.
5. Discrete point tests lend themselves particularly well to measuring formal English in an educational domain.

Discrete point direct tests (such as the BINL), were seen as a mixed blessing. On one hand, they involve, by definition, verbal interaction situations which are at least somewhat related to typical classroom interactions between student and teacher. On the other hand, however, they generally involve a higher level of training on the part of the tester and the scorer (particularly if they are the same person). The interviewer needs to be skilled in eliciting speech from the

respondent in relatively unstructured situations. This becomes very difficult with young children, especially when little time is available to establish rapport. Since the respondent's free responses must be analyzed for particular structures, vocabulary, etc., it is required that either the session must be tape-recorded and possibly even transcribed for later analysis or two people must be involved in the testing -- an interviewer and a scorer. Either of these alternatives is unattractive within the context of the present project with its staff of 100 or more interviewers (calculated at one interviewer present per interview) and a very few weeks to collect the data and score the criterion instruments. Thus, the discrete point direct approach was not given high priority.

Reduced redundancy tests were not prime candidates for two reasons: first, their validity as a global assessment of comprehension and speaking is somewhat controversial and, second, the dependence of these methods on respondents' reading and writing skills made them generally unacceptable.

This left two approaches, the discrete point indirect approach which has already been discussed, and the integrative, relatively more direct approach exemplified by the FSI Oral Interview. As applied to the present project, an integrative direct assessment would be one where the interviewer sets up a situation which would "call out" some of the skills necessary for performing adequately in an English-language classroom. Although no great amount of detail is known about exactly what those skills are, they clearly involve receptive and oral expression and receptive skills. Thus, the general sort of situation which suggests itself is one in which the interviewer engages the respondent in conversation and requests information, a narration, or statements of opinion. On the basis of that verbal interaction, then, the interviewer would rate the respondent on one or more scales of English proficiency. The advantages of this sort of procedure include its being more directly related to classroom interactions than are indirect discrete point

tests and quicker and easier to score than direct discrete point tests. Its chief disadvantage is that a good deal of interviewer skill may be called for, both in gaining the proper rapport with the respondent so as to obtain a representative sample of the respondent's verbal behavior, and in retaining an appropriate degree of objectivity in scoring to maintain reliability across a variety of social classes, ages, and ethnic groups. Clearly, the instructions given to the interviewer and his or her perception of this sort of task are crucial here. (An additional complication is that interviewers are generally trained to do everything in the interview strictly according to the manual both with respect to asking questions and recording responses. Thus, an activity such as this relatively unstructured one is often difficult for interviewers to do correctly.)

Given this preliminary review and discussion of the general approaches to English proficiency, Chapter III will describe the specific instrument development activities engaged in to produce both possible MELP instruments and the criterion measures which were then employed in the field test described in Chapter IV.

III. Instrument Development and Refinement

Activities related to the development and refinement of instruments began on June 1, 1975 and ended on July 18 when RTI held its initial training session for field test site supervisors. Most of the work was done in San Francisco, a site chosen for its varied ethnic populations and the relatively large numbers of limited English speakers. In particular, initial testing of possible instruments was done in the Latino, Chinese, and Filipino communities there. An additional consideration in locating in San Francisco was that CAL already had many civic and academic contacts in the area and thus could quickly recruit local personnel trained in linguistics and the social sciences to do the work.

A brief narrative of the principal activities which took place during this period can be found in Appendix 3.

During this phase, the staff organized itself into a number of overlapping teams, depending on the instrument to be developed and the ethnic group memberships of the team members. Since the time schedule was so short, instruments were constructed and tested in households, the data analyzed, and revisions implemented in a matter of days at most and sometimes in a matter of hours. Statistical analyses such as standard item analyses and correlations among scales within and across the three ethnic groups were done by hand and by using the Stanford University Computation Center. While these quantitative results were available and played some role in the development of the instruments, the largest factors in this phase of activities were the informal observations and intuitions of the staff and consultants who worked in San Francisco. As indicated in the Appendix, this group included both individuals with intimate knowledge of the ethnic groups and languages of interest and individuals with extensive experience in language testing, social

science research, and public education in San Francisco. It was only this unique blend of qualifications in the staff that made possible the production of some instruments in a five week period.

1. Development of Discrete Point Tests

The LESA - non-LESA distinction as legislatively defined appears to have three main foci: comprehension skills, speaking skills, and these as they are needed in instructional settings. Thus, it was desirable to address our discrete point tests to each of these.

Tests of comprehension. Existing comprehension tests have as a common property the following format: The interviewer pronounces a sentence or series of sentences and the respondent makes some sort of response from which it can be deduced that he "understood" the stimulus material. The response should be either non-verbal or minimally verbal so as not to confound comprehension with production skills. A common response is for the respondent to point to the one of several pictures that best illustrates the stimulus utterance. Knowledge of vocabulary and word order are particularly easy to index in this way. Another sort of receptive test is to give the respondent two sentences and he must indicate whether their meanings are the same or different.

Tests of Speaking. Many of these tests are available but nearly all of them tacitly assume that the respondent's comprehension skills are equal to or more advanced than his productive skills. Thus, they typically require the respondent to both understand and speak in order to correctly answer an item. Since these are discrete point items, each is focused on a particular linguistic feature or structure. A typical format is for the stimulus to include a sentence spoken by the interviewer, often a question, and usually referring to an object or picture which is present.

The respondent then must respond with an utterance that is both semantically appropriate and syntactically correct. The stimuli are designed so that the responses from native speakers will have a very high probability of containing the feature being tested.

Tests of Communication. Although both speaking and understanding of language are clearly called for in instructional settings, the ultimate requirement is that communication occurs between student and teacher. Thus, it was appropriate to look for a test that would involve some sort of overall communication task. Several of these exist or are under development. They usually involve some sort of task-oriented interaction between interviewer and respondent or among two or more respondents. The task is structured so that it cannot be accomplished without information being transmitted verbally, and it is easily determined when the solution has been reached. An example would be a two-person task where one has a set of blocks and the other a picture of how they are to be arranged. The object is for person 1 to duplicate the pattern in person 2's picture. While the relevance of such a task to everyday classroom communication requirements is arguable, it is a step toward forcing the respondent to use his linguistic skills in a communication context rather than in isolation.

Comprehension, production, and communication skills were thus the three principal foci of the test development effort, although other alternatives were pursued to some extent as discussed below.

There were several phases to the development activities. The first involved a massive search of all available materials on English Language Proficiency. From this set a number of tests were found which met many of the criteria of the project. This set was further scrutinized, then reduced, edited, and amended for pilot testing in San Francisco. The next phase involved changing or eliminating items

on individual tests based on pilot work with them in the Latino, Chinese, and Filipino communities in San Francisco. The LGRs' reactions to them also played an important role in this process (see Section III.6). During this phase whole tests were dropped from the battery. What emerged from these operations were two criterion batteries -- one for children and one for adults - which were then used in the field tests reported in Section IV.

Below we will present only the development of instruments which eventually found their way into the final tests; however, appended to this report is an account of our work with all instruments which we seriously considered and developed to some extent but which we did not include in the final tests. (Appendix 4)

The Oral Communication Test (OCT). This test was developed by Upshur (1971) and was used in the present study to test communication skills of children and adults. It is an individually administered test for adults of ability to communicate in a foreign language, and had been used with respondents as young as 10 years old. The test contains thirty-six communication tasks.

Upshur (1971) describes the tasks as follows:

> (1) The examinee is presented with four pictures differing significantly on one or two conceptual dimensions. These (pictures) may represent, for example, a person performing four different 'actions', or the four conjunctive possibilities of a man with or without a hat walking up or down a staircase.
>
> (2) The examinee is instructed to provide a single sentence description to a visually remote audience of one picture which is randomly selected from the set.
>
> (3) The audience -- who is the examiner -- makes a best guess as to which picture is being described.
>
> (4) The examinee's directed intentions (about which picture to indicate) are compared with the examiner's guesses (1971:438).

The test yields two scores: The number of messages successfully communicated, and time required for communication.

Respondents are first given oral instructions and four unscored, example tasks. If they are unable to perform two of the last three examples, testing is not continued. Each subject is presented with a key in the form of a list numbered from 1 through 36. Following each of these numbers is a letter: A, B, C, or D. These letters refer to the one picture in the four picture set which the subject is to identify by his utterance. Different keys are used; in each key the pictures indicated have been randomly selected in order that the examiner cannot learn which pictures a subject is attempting to indicate.

The stimulus pictures measure two and one-half by two and one-half inches. These are aligned horizontally on a card measuring six by twelve and one-half inches. In the upper right corner of the card is the number of the test task: 1-36. Below the four pictures are the letters A-D reading from left to right. The thirty-six test cards and four example cards are placed before the respondent in a stack face up. The respondent's key is placed facing him and closer to him than the picture cards.

When the respondent is ready to attempt an item he refers to his key and turns over the currently exposed card in order to reveal the item he will attempt to communicate. He is given three seconds to examine the set in order to see the significant differences among the four pictures. Then the examiner gives him a cue to respond, saying either, "Describe the correct picture," or, "What is the man doing?" As soon as the cue is given the examiner begins timing the respondent with a stop watch. Timing is stopped as soon as the respondent has completed his single sentence description, or at the end of twenty seconds if the examiner records his guess of the keyed picture for each item according to the respondent's utterance. No attempt is made to evaluate linguistic aspects of a respondent's speech.

After the test session, the examiner compares the respondent's key with his own recorded guesses. The number of corresponding numbers is the respondent's message score. The total time used in responding to the thirty-six items is the time score.

The following modifications to the test were made during the San Francisco pilot work. No time limits were set -- the subject could look at the stimulus for an unspecified length of time before he responded. He could take as long as he wanted to respond. This modification was made because it was felt that a time restriction might penalize Navajo speakers who reportedly have long latencies in conversations as a normal characteristic.

All communication tasks were arranged in a booklet. For each task an "X" was put below the stimulus to be described. There were four different sets of materials: all contained the same items but differed in terms of the specific picture in each item to be described. As mentioned before this was done so that the examiner would not become familiar with the stimuli and memorize the sequence of correct answers.

Other amendments were also made as a redult of field experience. The number of communication tasks was eventually reduced from 36 to 15 and all pictures were redrawn to make them more realistic. Although time scores were taken, they were not used for the final analysis. Otherwise the scoring procedure was the same as that described by Upshur.

The Adult Production Test (APT) was adapted from the Ilyin Oral Interview procedure (Ilyin, 1972). The test was developed to test an adult ESL speaker's oral proficiency in English. In the original procedure, the respondent is shown a picture and asked a question to elicit a specific grammatical structure. There were 50 items in the test. Each response could receive a maximum of 4 points: 1 for information, 1 for word order, 1 for verb structure, and 1 for other grammatical elements.

In the first phase of the San Francisco testing the following modifications were made. The test was given to adults and children, and the instructions were simplified. Thirty of the original items were used. These items had been specified by Ilyin (personal communication) as being the most discriminating.

The instrument was further modified during the pilot activities. It was too difficult for children and thus only given to adults. The items were further reduced to 16. All pictures were redrawn to make them more realistic. The scoring procedure was simplified. Each response could receive a maximum of two points: one point for correct information, and one additional point if the grammatical structure of the response was correct as well. Also, after failures on five consecutive items, the test was discontinued for that respondent.

The Adult Comprehension Test (ACT) was based on the items of the CELT (Upshur, et al, 1964). The CELT was developed to test English Proficiency in adult speakers of ESL. Our interest was in the Listening section of the test which is composed of two parts. In part 1 the subject hears a question and then has to select from four written alternatives the best response. For example the respondent hears When are you going to New York? and then reads the following alternative answers:

a) to visit my brother

b) by plane

c) next Friday

d) I am

He then marks the most appropriate one. There are 20 such items. Part 2 is composed of 20 items. Here the respondent hears a sentence such as George has just returned home from vacation and then reads four alternative sentences:

a) George is spending his vacation at home.

b) George has just finished his vacation.

c) George is just about to begin his vacation.

d) George has decided not to take a vacation.

He is asked to mark the sentence which is closest in meaning to the one he has heard.

The basic idea behind the test was intriguing even though the form had to be greatly changed because a paper and pencil test was undesirable. As modified by CAL, part 1 required the examiner to ask a question. He then orally gave the respondent two different answers. The respondent had to indicate which one was best. In part 2 the examiner said two sentences. The respondent was asked to indicate whether they were the same or different in terms of meaning.

Since time pressures dictated a speedy start in testing and revising this instrument for use in the field test, the necessity of negotiating with the publisher for permission to make modifications was circumvented by simply using the general logic and format of the items but entirely recreating the test ourselves with all new items. Even so, of course, many of the same language structures were tested as are tested in the CELT.

There were 30 question and answer items and 43 sentence pairs. Both children and adults were given the test. By the end of the San Francisco phase the following modifications were made.

a) The Question-Answer section was totally eliminated. Examiners reported that the task was too difficult. One of the major reasons for this seems to be that there was no context for these questions.

b) The task was too difficult for children. It was only given to adults.

c) The final number of items was reduced from 43 to 10. The 10 surviving items were selected on the basis of having high part-whole correlations with the total score of the 43 items. The resulting instrument was called the ACT or Adult Comprehension Test.

MAT-SEA-CAL. This test was developed by Joseph Matluck and Betty Mace-Matluck (1975) under the auspices of the Seattle Public School Board and the Center for Applied Linguistics. It was developed to measure the child's ability to understand and produce distinctive characteristics of spoken English. It was originally intended for children in Kindergarten through Grade 4. CAL adapted sections of this test which were eventually used to measure English receptive and production abilities in children.

Part 1 of the original test had 27 items. For items 1-17, the examiner says a sentence and the child points to one of four pictures which best gives its meaning. In items 18-27 the examiner gives a command (e.g., Stand up) to which the child responds.

In the pilot work, the commands were eliminated from this section because they were too easy and thus did not discriminate between good and poor proficiencies -- only between poor and no proficiencies. Minor modifications were made throughout the pilot-test to items 1-17, and the final instrument was composed of 12 items derived from the original ones. The pictures were redrawn to make them more realistic and the number of alternatives in each item was reduced to three. As in the other tests described, administration was terminated after 5 consecutive failures.

Part 3* of the Mat-SEA-CAL is called "structured response" and is meant to test oral production. The task is very similar to the Ilyin Oral Interview described above. The respondent is shown a picture and asked a question about it. The question is so designed to elicit a specific grammatical structure from the subject. There were 28 items in the original MAT-SEA-CAL, each worth one point if the response was grammatically correct.

---

* Part 2 is an imitation task. It was never considered in that an imitation procedure was built into the ETS test discussed in the appendix. Results of that test indicated considerable difficulties in scoring an imitation test; thus, even when the ETS test was dropped, the MAT-SEA-CAL imitation section was not considered.

In the pilot work, the test was given to children up to 14 years much as described above. The following modifications were incorporated into the final items.

1) 20 of the 28 items were retained

2) the pictures were redrawn

3) the scoring procedure was changed. Each answer was given one point for correct information, and one additional point for being grammatically correct in addition.

To summarize, the following table shows which tests were used in the final battery, to whom they were given, and what each was meant to measure. All tests were discrete point and indirect in their general approaches to the measurement of language proficiency.

Name of Subtest	Measures	No. of Items	Possible points
Adults			
1. Adult Comprehension Test (ACT)	Reception	10	10
2. Adult Production Test (APT)	Production	16	32
3. Oral Communication Test (OCT)	Communication	15	15
Total		41	57
Children			
1. MAT-SEA-CAL-I	Reception	12	12
2. MAT-SEA-CAL-II	Production	20	40
3. Oral Communication Test (OCT)	Communication	15	15
Total		47	67

The developing and final forms of these tests are reproduced in Appendices 9 and 10 respectively.

2. Development of the Direct Observation Rating Procedure (DORP)*

There were two motivations for developing this measure:

1. To serve as a criterion measure which would qualify as a direct measure of English proficiency based on face-to-face interaction and observation.

2. To provide a "back-up" MELP instrument in case none of the opinionnaire-type questions were satisfactorily predictive of the criterion measures.

Since the constraints of the project dictated that it must be administered by an interviewer (rather than a teacher) in the household (rather than in a school), there were severe limitations on just how direct a measure the DORP could be. One way in which directness could be preserved was to develop the descriptors of the scale positions with the help of teachers rather than linguists or researchers. Teachers were also consulted in the formulation of the speech elicitation situation.

Procedure: The development of the DORPs for children and adults were developed separately but in parallel. In both cases, several steps were involved.

1. Elicitation and recording of free-speech, both conversation and narration from respondents of various ages, linguistic backgrounds, and English proficiency.

2. Elicitation from teachers of ratings of the speech samples plus comments on the properties of the samples that determined their ratings.

3. Compilation of these data into descriptions of a graduated scale of English proficiency.

a. Elicitation of speech samples: In the course of data collection involved in refining other instruments, recordings were made of brief conversations between interviewer and respondent. The respondents were asked a range of open-

---

* Special thanks go to Amador Bustos, Carolyn Karelitz and William Sinclair for their contributions to the development of this measure.

ended questions such as "What is the most exciting thing that ever happened to you?" "What is your favorite TV program?" "Tell me about your best friend." etc. Respondents were then shown a book of photographs, asked to describe several photos, and asked to tell what they thought was happening in each picture. Such data were collected from 15 children and 8 adults. The children ranged in age from 6 to 13 and included Latinos, Chinese, and Filipinos. The ages of the adults ranged from 18 to 70 with all three ethnic groups represented. The speech samples were then copied onto two master tapes, one for children and one for adults.

b. Judgments of speech samples by teachers: Two sets of teachers were employed to judge the speech samples. The 24 teachers judging the children's tape were all certified, employed elementary school teachers in the Bay area. All had had experience with children whose native language was not English. Fourteen teachers judged the adult samples. They were all actively teaching in adult education programs in the Bay area. All teachers made their ratings in groups of from six to 14 people. The procedure was as follows:

1. The need to develop a direct observation scale was explained.

2. Each teacher was provided with a form on which to rate each sample and write comments about it. (See figure 1). They were to use a seven-step rating scale.

3. Before hearing each sample the teachers were told the age of the person whose speech was to be heard.

4. The first two samples to be heard were the least proficient and most proficient of the group as judged by the project staff. The teachers were told that they were to rate them as 1 and 7 respectively.

5. As each sample was played, teachers were asked to make their ratings and then to write as completely as possible the reasons why they rated the

speaker as they did, paying special attention to specific features that they had noted in their experience as being predictive of academic success or failure in a non-native English speaker.

6. After the samples had all been played, a group discussion was initiated about language requirements for success in the classroom.

7. The session lasted two to three hours overall and each teacher was paid $25 for participating in it.

c. Analysis of responses: The data analysis was essentially the same for adults and children. First, the mean rating and its associated standard deviation were computed for each sample of speech. Speech samples eliciting widely divergent ratings from the teachers (as evidenced by high standard deviations) were eliminated from further consideration. A list was then made of all the teachers' descriptive comments for the samples remaining at each step of the scale and a content analysis was made of the comments about the samples in each step. The comments were categorized with respect to the following aspects of speech behavior:

1. Fluency: hesitancy or quickness of response, need for prompting.

2. Comprehension: comprehension of questions and instructions, of sequences, of events, ability to draw inferences.

3. Sentence Structure: Complexity of sentences, word order, use of prepositions, articles, and verb tenses, variety of sentence types.

4. Vocabulary: Use of adjectives, slang, words from the native language, and colloquialisms.

5. Pronunciation: Interference, intonation, accent.

Next, a seven column (mean rating positions) by five row (dimensions of language evaluation) matrix was constructed. Each cell contained all comments about all samples occupying that particular scale position dealing with that particular

dimension of evaluation. Separate matrices were constructed for adults and children. Inspection of those matrices immediately indicated that there was no apparent difference between the descriptors of positions 2 and 3 on one hand and 5 and 6 on the other. Thus, the scale was collapsed to a 5 point scale. Finally, the most frequent comments in each of the cells were combined into several sentences emphasizing the distinctions between neighboring cells. The choice was then made to eliminate the five separate dimensions from the final DORP scale since the instrument had ultimately to yield a single rating for each respondent. Descriptions of the five global scale positions were synthesized from the columns of the matrix. Those descriptions were the ones provided to the interviewers and are reproduced in Appendix 12.

d. The Elicitation Situation: The final aspect of the DORP to be defined was the elicitation of the speech sample. This was a significant problem because of the requirement that the situations be at least somewhat standardized over the entire range of ages and ethnic groups. The general problem of obtaining useful spontaneous language samples is well known by sociolinguists, and there are apparently no easy solutions (cf. Wolfram and Fasold, 1974) even under the best of conditions. It amounted to finding situations in which people with very different backgrounds and interests would all talk with equal ease and volubility. Unfortunately, even if that objective were achievable, we had no time to test various procedures. Thus, the solution adopted was merely to have the interviewers ask three open ended questions of each respondent with further instructions to add to those questions in any way that would be likely to get the respondent talking. The questions were picked from among those that seemed most effective when eliciting the speech samples used in the development of the rating scale. They are included below.

ADULT QUESTIONS

1. "Could you take a second and think of the one person who has made a big impression on you, and tell me, as much as you can about that person. (pause) I'll just listen, and you tell me. Take your time."

2. "Now if you will, I'd like you to think back to one of the most exciting experiences in your life. Tell me as much as you can about that experience."

3. "Now a final question. Take a second to think about this question. If you could do anything you wanted to do today, what do you think you might do? Tell me as much as you can about what you might do."

CHILD QUESTIONS

1. "Could you take a second and think of your best friend, and tell me as much as you can about that person. (pause) I'll just listen and you tell me. Take your time."

2. "Now if you will, I'd like you to think back to one of the most exciting places that you've been to. Tell me as much as you can about that place."

3. "Now a final question. Take a second to think about this question. If you could do anything you wanted to do today, what do you think you might do? Tell me as much as you can about what you might do."

Figure 1: Form used by teachers to rate and comment on speech samples

SAMPLE # _____

Please listen carefully and make any notes in the space provided below:

NOTES:_____

_____

_____

_____

_____

_____

_____

(If more space needed, please write on back of sheet.)

Please rate the sample on the basis of the child's likelihood of succeeding in (or benefiting from) a monolingual English class (circle one).

1 2 3 4 5 6 7

(least likely) (most likely)

Give as many reasons as you can for rating this sample the way you did:

_____

_____

_____

_____

_____

_____

_____

_____

3. The Monitoring System

In CAL's proposal to NCES, the need was expressed to develop an objective behavior monitoring system to obtain data on the nature of the interactions between interviewer and respondent during the asking and answering of MELP questions. This was seen to be particularly important because of the possible cultural and linguistic differences between monolingual English speaking interviewers and potential LESA individuals. (It was not at all clear at the time in what numbers Census would be able to hire interviewers who were members of the ethnic-linguistic groups involved.) CAL planned to have its staff members monitor the RTI interviews to collect both objective and impressionistic data on strengths and weaknesses of the questionnaire and procedure. Without exception, these monitors were members of the research staff who had developed the MELP questions, the test, and the DORP in San Francisco and had conducted many such interviews themselves, thus they were well-acquainted with the objectives of the project and the intended uses of the instruments.

In mid-June, Dr. Jeanne Freeman was given the assignment of developing an objective behavior coding system to monitor the interaction in interviews. The remainder of this section is her report of the development activities.

The development work began with an extensive review of the literature on interaction analysis systems (e.g., Simon and Boyer, 1967; Rosenshine and Furst, 1971; and Dunkin and Biddle, 1974) and the literature on non-verbal communication (e.g., Mehrabian, 1972). This review of the literature, coupled with consultation with Dr. Jere Brophy of the University of Texas at Austin led to the selection and adaptation of verbal and non-verbal categories from already existing systems and the development of categories appropriate for this specific study.

A preliminary set of categories was developed for verbal and non-verbal behaviors. The non-verbal categories reflected major areas: proxemics (distance),

haptics (touching), kinesics (body movements), oculesics (eye behaviors). In addition, verbal categories were developed to differentiate and record various phases of the interview. This initial list of behavioral categories was submitted to the development staff (who represented various ethnic groups) in San Francisco. The staff rated the categories in terms of appropriateness for the different ethnic-linguistic groups. Although there were several categories that were questionable, the first draft of the monitor's interaction analysis system was developed, including definitions and examples of the categories.

This first system divided the interview into three sections: the introductory/orientation phase, the questioning/answering phase, and the closing phase. Each phase contained categories specific to that phase (i.e., in the introductory/orientation phase, specific verbal and non-verbal greeting behaviors; in the closing phase, specific verbal and non-verbal leave taking behaviors). However, each phase was also coded according to a single set of global rating scales developed to assess high inference behaviors, such as responsiveness and tension.

The first set of categories for the introductory/orientation phase of the interview included verbal greeting behaviors, such as exchange of pleasantries and receptive-unreceptive comments, and non-verbal behaviors, such as distance from interviewer, touching behaviors, and facing the interviewer. The global rating scale coded at the end of this phase and at the end of each subsequent phase included five-point rating scales representing general behaviors (pleasant-unpleasant, responsive-unresponsive, tense-relaxed, tolerant-intolerant, open-withdrawn, formal-informal).

The categories for the question/answer sequence, in which the interviewer asked the census-type questions and the criterion measures, included four five-point rating scales (willingness to respond, nervous-calm, brief-detailed, positive-negative) to be completed for each item. Toward the close of the question-answer

sequence, the monitor rated the respondent according to the occurrence of specific non-verbal behaviors, (e.g., facial expressions, facing the interviewer, looking toward the interviewer, leaning toward the interviewer, stiff posture, tense hand/leg movements).

The categories for the closing phase of the interview included verbal leave-taking behaviors, such as exchanges of personal information and comments to maintain or close the interaction, and non-verbal behaviors, such as walking the person to the door, distance from the interviewer, and touching behaviors. After recording these behaviors, the monitor would code the respondent's behavior according to the same global rating scales; however, in this phase, the monitor recorded changes in global behaviors. For example, the monitor would check pleasant-unpleasant for one of the following: a mixed pleasant/unpleasant response, a change from pleasant to unpleasant, a change from unpleasant to pleasant, or no change. Therefore, the monitor could infer general characteristics of the respondents' behavior and record the general pattern of the entire interview for each global category.

The first version underwent modification with the help of three CAL research assistants* in San Francisco and resulted in a considerably simplified category system: (1) the specific verbal and non-verbal greeting and leave taking behavior categories in phases 1 and 3 were eliminated, and the list of non-verbal behaviors in phase 2 were substituted. (2) the specific non-verbal categories and the global rating scales were collapsed somewhat. For example, rather than having separate categories for nervous hand movements, nervous arm movements, nervous leg movements, or nervous foot movements, these were collapsed into a category nervous hand/arm/leg/foot movements. Also, since eye contact was so variable among ethnic groups,

---

* Evangeline Kamitsuka, Michael SamVargas, and Richard Chambers

the various eye contact behaviors were deleted and incorporated into a category looking toward the interviewer. Pleasant and friendly were collapsed into one category. These changes and modifications constituted the second draft of the coding system.

After arriving at the set of categories for the second version of the manual, the research assistants refined the categories in the system by elaborating the definitions and examples and by reducing the five-point scales to three-point scales. During this phase of the development, the objectives of the monitoring system were reassessed and to some extent reformulated. The objective of assessing the validity of the respondent's answer remained; however, the objective of assessing affective verbal and non-verbal interactions was considered of secondary importance; therefore, the categories were redesigned to focus strictly on the respondent's answers to the MELP questionnaire and whether the interviewer achieved the objective of the question (i.e., obtained the information called for by the question). The second phase of the interview, the question/answer phase became the basic framework for the revised version in which several categories were coded for each question/answer unit.

The third version of the monitoring system involved structuring and elaborating the question/answer phase in which each question answer unit would be coded according to several categories. In the question/answer sequence, response and detail remained as categories. In addition, several categories were added (other answers, relevant answer, seek clarification, rephrase and achieve objective). The framework for these categories consisted of four three-point rating scales (response, detail, nervous, and attentive) and five checklist categories (relevant answer, rephrase, seek clarification, another answer , and achieve objective). After developing definitions and coding procedures for these categories, the

staff practiced coding with this system. Preliminary testing led to further changes: (1) deletion of the high inference categories (nervous and attentive) (2) changing all categories except detail to checklist form, and (3) including cases in which the interviewer or respondent uses his native language. Also, to facilitate the actual coding, the categories were logically structured into the following superordinate categories:

Problems of the respondent in making a response to a question

-- Does not respond

-- Answers with information irrelevant to the question

-- Another person answers the question

-- Respondent seeks clarification

-- Respondent uses language other than English

Interviewer Behavior

-- Interviewer rephrases the question with or without an explicit request from the respondent to do so.

-- Interviewer uses language other than English

General

-- The objective of the question appears to have been achieved

-- Amount of detail of information given by respondent in answering the question (insufficient, sufficient but minimal, more than sufficient)

These categories were selected to code only what the interviewer or respondent said in English; questions or answers in translation were coded only as uses other language. In order to standardize the monitoring, this procedure was required due to the variability of the monitors, some of whom did not speak the language of the ethnic-linguistic groups.

The final categories were incorporated into coding sheets designed to identify each census question by a number and code word, so the monitor could readily identify the answers to each of the questions. For example, the monitoring form corresponding to MELP question #1 (date of birth) was:

#1 Date
No response _____
Irrelevant answer _____
Another answer _____
Seeks clarification _____
Rephrase _____
Int. other language _____
Resp. other language _____
Achieve objective _____
Detail 1 2 3

In addition to coding each question-answer unit for the census questions, monitors recorded comments about specific unusual occurrences, such as the respondent not completing the interview, the respondent having auditory problems, the respondent having difficulty reading flash cards, or the respondent being resistant or inattentive. Also, monitors recorded any other circumstances that may have affected the respondent's performance or would affect interpretation of the data.

Preliminary coding to establish interjudge reliability was done by Freeman, Kamitsuka, SamVargas, and Chambers. Major disagreements on problems of definition were resolved before establishing interjudge reliability on each category. Reliability data for each category were based on the percentage derived from the formula:

$$\frac{\text{\# unanimous agreements among judges (4)}}{\text{\# occurrences of the category}}$$

For all categories, 80% agreement or above was established. It was felt these percentages were sufficiently high to justify use of the system for the field test.

A final draft of the category system was developed for use in training the other staff members to use the system. Training included general overview and discussion of the categories and practice coding using videotapes. Results of the reliability assessments were fed back to the participant coders and discussed. Training was completed before the staff left San Francisco for the various field test sites. A copy of the manual is appended to this report. (Appendix 13)

4. Development of the MELP Questions

Among the published accounts of using questionnaires to collect data on language proficiency and use patterns (e.g., Lieberson, 1966; Mackey, 1966; Kelly, 1969; Harrison, Prator, and Tucker, 1975; and Committee on Irish Language Attitudes Research, 1975), the one relied on most heavily in the present project was that by Fishman, Cooper, and Ma (1971). (Both Fishman and Cooper served as occasional consultants on the project.) Generally, this literature indicated that individuals can rate their own language proficiency fairly accurately (as compared with their performance on tests), and that both their current use of the language and their educational history involving the language correlate quite highly with test scores as well. Thus, the initial foci of the MELP questions were five-fold:

A. Screening Questions. In chapter I of this report, the need for a set of screening questions was discussed. They were to define the pool of potential LESA individuals as characterized by PL 93-380. In particular, they were to determine:

a. Place of birth
b. Usual language spoken by the individual
c. Usual language spoken by the individual's household
d. Parents' usual language (for children)

B. Self-rating Questions. These were questions asking the respondent to directly evaluate his own ability to speak and understand English. Respondents were also asked to rate their proficiency in their non-English language on the possibility that proficiency in one language might be inversely related to proficiency in the other. Proxy respondents were asked to rate another person in their household.

C. Language Use Questions. Assuming that proficiency in a language is directly related to the extent and variety of its use in various situations, a number of questions were tested which explored the respondent's usual language in the home, at school, at work, and with peers.

D. Educational History. Since the LESA concept is defined relative to educational settings, questions were created dealing with:

(1) number of years of formal education completed
(2) country in which the education was received
(3) number of years in which English was the principal language of instruction
(4) whether the individual had ever been informed by a school official that his English was insufficient for educational purposes
(5) whether the individual had ever been held back in school (because of deficiency in English)
(6) whether he or she had ever participated in a bilingual program
(7) whether he or she had been enrolled in school in the last year

E. Mass Media Questions: Several questions relative to the respondent's use of various English language mass media were explored on the hypothesis that the regular use of English mass media would imply proficiency in English. The converse, of course, would not be a reasonable implication (i.e., that one not using mass media was not proficient in the language).

Procedure: The procedure used in developing and testing these questions was as follows: Dr. Terry Webb and Dr. Alberto Rey were principally involved in producing drafts of the MELP questions. They were closely guided by Leslie Silverman of NCES while he was on site. The questions went through so many editions that it is not useful to try to trace their evolutions in detail here; however, several sequential versions of the questionnaire are appended to this report. Generally, the procedure was as follows:

1. An edition of the questionnaire was produced and distributed to the various teams developing the tests.

2. They would use the questionnaire for one or two days of interviewing in the Latino, Chinese, and Filipino communities.

3. A meeting of the entire staff would be held in the late afternoon and the experience with each question in each ethnic group would be discussed in detail.

4. Revisions would be made over night and a new version typed, reproduced and distributed to the teams by noon the next day.

5. Etc.

Since the questions in the NCES "Survey of Languages" had already gone to press as part of the July, 1975 edition of the Current Population Survey, they were generally included in a form unchanged from the CPS. This would enable some comparisons of their adequacies relative to some created by the CAL staff which covered approximately the same topics.

Finally, on July 12, the then current version of the questionnaire was reproduced for distribution at the July 13-14 meeting of the LGRs. That edition is appended to this report. (Appendix 14)

5. The Language Group Representatives

Selection. The nature and purpose of the LGR advisory committees demanded that they be composed of individuals who were members of or who had worked with the various linguistically different groups in the United States. Emphasis was placed to identify and select individuals involved in community work on the political, social and/or religious levels. Similarly, attention was placed on the selection of participants who had had a chance to work in areas where the concept of education had been actively discussed or been a major goal.

Due to the linguistically heterogeneous nature of the American populace, CAL felt that a number of language groups had to be represented. Consequently, five major language groups were identified with subgroups within each. The five major language groups were Spanish, Chinese, East Asian/Pacific, Native American, and European/Near Eastern. Equally important was that areas of the country where the language groups were found should also be represented--the rationale being that a language group in one part of the country did not necessarily have the same background, goals, desires, needs and degree of English language proficiency as a similar group in another part of the country. For example, Chicanos in Texas, tend to be located more in rural areas and have perhaps more ties to the Spanish language a and culture than their counterparts in the Midwest. For this reason, a relatively large language advisory committee was assembled. Consequently, advisors were drawn from (1) specific dialects/languages within each of these language groups and from (2) various areas of the country where these languages/dialects were represented.

The suggested plan called for a representative group of Spanish-speaking Mexican Americans from the West Coast, Texas, and the Mid West; Puerto Ricans from the East Coast and Chicago; and another group from the Cuban, Dominican, and Central American communities. Organizations like L.U.L.A.C., National Task Force de la

Raza, El Congreso, Mexican American Council on Education, United Migrant League, ASPIRA, Puerto Rican Legal Defense Fund, and other Spanish speaking community-based and/or -minded organizations served as sources or contacts for this advisory board.

In addition, an advisory board was selected to incorporate the Chinese perspective. Representatives from West Coast, East Coast and Chicago community organizations were invited to assist in the tasks for this board. Likewise, representatives from the East Asian/Pacific language groups were identified and involved. The Korean, Vietnamese, Japanese, Filipino and Samoan communities were canvassed for advisory board representation.

The Native American advisory board was made up of a representative group of Navajo, Sioux, Mikasuki/Seminole, Papago, and Eskimo, as well as representation from the Northwestern tribes. Organizations like the National Congress of American Indians, National Indian Education Association, United Sioux Tribes, United Southeastern Tribes, United Indians of All Tribes Foundation, and the Navajo Division of Education were identified as sources or contacts for this board.

Finally, the European/Near Eastern perspective was incorporated by including representatives from the French (New England, Louisiana, Haitian), Italian (East Coast), Portuguese (New England), Greek, Polish (Chicago), Serbo-Croatian and Arabic (Detroit) language communities.

The above national groups reflected an approximate total of 45 individuals who were invited to form the advisory committees. The geographical areas of concentration which were identified were in no way fixed; rather, these were areas which, based on current census data, seemed to have a significant number of the aforementioned population groups.

6. Role in Instrument Development and Use

LGR Meeting #1. The first LGRs were scheduled for meetings June 10 and 11. Subsequent groups came to CAL offices in Arlington, Virginia every two days until June 18-19 (Spanish, Native American, Chinese, Asian/Pacific and European).

The morning of the first day was spent introducing the LGRs to CAL and its work and NCES and its work. CAL's involvement in the NCES project was outlined carefully. Moreover, the project was placed in perspective relative to current legislated mandates. Likewise, the project was discussed at length to insure that the LGR's understood what the consequent MELP would do and not do, and the purposes of its use.

The afternoon session was devoted to several points of discussion. First, the concept "instructional/educational difficulty" (quoted from current legislation regarding bilingual education) was introduced, and attempts were made to arrive at a group definition. Then, several reports were given which focused on past and current language assessment in the represented LGR communities.

The second day was devoted to a review of current research regarding theory and practice in language testing. This was supplemented by a review of effective and tested sociolinguistic field methods. Discussion focused on the consequences of "mistakes" in data gathering.

Criterion and candidate MELP measures for language assessment were then introduced and discussed. It was pointed out that project items or measures could not be of a criterion type, rather, they had to follow "census type" questions. Nevertheless LGRs were asked to consider the initial battery of criterion measures and assess them for their face validity. Finally, the LGRs were given an opportunity to make recommendations regarding potential cultural and linguistic biases in the

MELP format and items (for those proposed for the initial field testing). Likewise, recommendations were accepted regarding current, sensitive guidelines to be followed in order to facilitate all data collection.

Every LGR meeting followed basically the same agenda and content. Gil Garcia, Leann Parker, Dr. William Leap, Diana Riehl, and Dr. Roger Shuy collaborated in these efforts. (See Appendix 15 for LGR reports)

LGR Meeting #2. Although the LGRs made preliminary comments about the kinds of instruments that would be appropriate for their respective groups (both MELP questions and criterion instruments) during their initial meetings in June, their main opportunity for concrete input to the project came during Meeting #2 in San Francisco on July 13-14. Upon arrival they were given packets containing all of the instruments developed in the pilot activities (discussed in the preceding sections of this chapter). The first morning was spent in a general briefing by Walter Stolz on the activities to date, the design of the field test, and a review of the general objectives of the MELP project and the SIE. Then Earl Gerson of the Bureau of the Census briefed the group on the general sampling plan to be used in the SIE.

In the afternoon, the CAL staff acquainted the LGRs with the instruments and the general interviewing procedures to be used in the field test. This was done by role-playing interviews using the LGRs as respondents. Video tapes of several interviews made in the last days of the pilot work were also shown.

During the remainder of the conference intensive discussions were held within each area group of LGRs relative to specific aspects of the instruments which should be modified or eliminated. Each representative was asked to submit an individual critique of all materials; however, each group also prepared a single report to be presented to the conference as a whole. These reports were presented and discussed on the last afternoon (see Appendix 15). As can readily be seen they range from comments on individual items to critiques of the government's philosophy toward bilingualism and bilingual education.

During the ten days between the LGR meeting and the beginning of the field testing in Miami and El Paso, both MELP questions and criterion instruments underwent considerable change. The MELP questions were revised in group session by Stolz, Webb, and Troike of CAL, Horvitz and Weeks of RTI, and Dr. Dorothy Waggoner of NCES. The final field test questions are reproduced as Figure 1 in Chapter V of this report. The tests were revised by Strick in cooperation with the RTI graphics department. They are appended to the report.* Some specific changes in the instruments stemming from the LGR's input were:

1. The MELP Questions

   a. Some questions were included to probe the reespondent's knowledge of his first language as well as his knowledge of English (e.g., questions 9, 10, 11, 15).

   b. On questions calling for a proficiency rating, the negative connotations of the lower steps were removed.

   c. Question 4 was changed in accord with a suggestion from the Chinese group.

   d. Questions were asked separately about newspapers, magazines and books.

   e. A question about the language used at work was included, as well as some questions about type of work.

   f. Several questions were removed which seemed to have little to do with English proficiency.

2. The Adult Production Test (Illyin)

   a. All pictures were redrawn to make them look more professional.

   b. The beach scene was eliminated, and a scene in a park was substituted.

3. The Mat-Sea-Cal

   a. All pictures were redrawn

   b. An item involving a monkey climbing a tree was eliminated, and another item was substituted ("It's on the corner").

* See Appendix 9

4. The OCT

   a. Stick figures were redrawn more realistically.

   b. The administration procedure was simplified.

5. The ACT

   a. An additional example was incorporated into the instructions.

LGR suggestions about interviewing personnel were followed by hiring approximately one-half of the interviewers in each site from the ethnic group being surveyed. Also, a more thorough orientation-training program was carried out for each site lasting three days instead of two as originally planned. In training interviewers for the Navajo site, Dr. Robert Young from the University of New Mexico, was brought in for two days to provide a general orientation to Navajo culture.

A concern about speed of responding to the tests was expressed by the Native Americans in particular. They thought that many Navajos may require more than the usual time limit if 10 or 20 seconds per item to respond with the correct answer. Thus, the interviewers were instructed to allow as much time as the respondent needed to give an answer.

Site Visits by LGRs. Several LGRs monitored the field test activities in the various sites. They traveled with one or more interviewers on their rounds and then made a report to the RTI supervisor and the CAL monitors. Suggestions for changes in procedure were referred to RTI's and CAL's central offices. LGR visitations included:

Miami

Willy Gort
D. G. Kousoulas

Arizona

Dillon Platero
Fidel Davila

San Francisco

Ling Chi Wang
Danilo Begonia

No LGRs visited El Paso because the work had ended there before a schedule could be set up. Dr. Robert Young spent a day monitoring interviews in Arizona as an expert in Navajo culture. Each of the above LGRs reported back to their respective groups at LGR Meeting #3. (See Appendix 15 Reports.)

Meeting #3. The third LGR meeting was held in Arlington on September 3-4, 1975. The main purpose of the meeting was to brief the representatives on the field test procedures and preliminary results and to obtain general suggestions with respect to analyses and interpretations of the data. The proceedings of that meeting are appended to this report.

At the time the meeting was held, virtually complete data from Miami and El Paso were in the computer; however, only about one-third of the data from the other two sites had been processed into computerized form. Using the data available, frequency distributions and crosstabulations of MELP questions vs. test scores were constructed and distributed to the representatives. Stolz explained this material and discussion both in the plenary session and in groups ensued about how these results would be used to produce a MELP instrument and how that instrument would be used to categorize people as either LESA or non-LESA. Summaries of these discussions may be found in the proceedings. (See Appendix 15 Reports.)

IV. Field Testing the Instruments

1. The Basic Design

A principal step in the development of any instrument is the field testing phase. In a field test, the instrument is used in a context as close as possible to that in which it will eventually be employed in the survey proper, but additional data are also collected which allow for an evaluation of the trial instrument's performance. The most important evaluation which could be made in a case such as the present one is concurrent validity, and that was in fact the primary objective here. Concurrent validity was evaluated by correlating the items in the trial MELP instrument with several "criterion" measures of English proficiency. The development of two such instruments, the test and the DORP, have already been described in detail. The obvious way of obtaining correlations of MELP items and criterion measures is simply to collect all measures in a single interview and then compute correlations for all possible pairs of these variables. This was what was done with the MELP items and the test and DORP using a concurrent measurement validation design.

When the criterion variable is not continuous but rather categorically defined, a known groups validation design is possible. In this design, respondents are chosen for participation in the study on the basis of their having been identified as belonging to one or another category of the criterion variable before the field test instrument (the MELP) is administered. A known groups design was possible in the present study because school systems serving populations that include considerable numbers of children with native languages other than English screen such students for participation in special English-as-a-second-language or bilingual education curricula. Such screening procedures constitute local operational definitions of the concepts LESA and non-LESA in the sense that "passing" such a screening

procedure is taken by the school as evidence that the child can succeed in a standard monolingual English instructional environment, i.e., he is not LESA. Conversely, if the results of the screening procedure suggest the advisability of enrolling the student in special programs, this is equivalent to indicating that the student might encounter some "instructional difficulty" in the regular curriculum, i.e., he is LESA. To the extent that such screening procedures are well-constructed for their purpose they produce appropriate known groups against which the MELP can be validated. They are particularly valuable in that they provide a non-arbitrary cutting point between LESAs and non-LESAs on the continuum of English proficiency non-arbitrary because the cutting point is implicitly referenced against the school's curriculum.

The disadvantages of using the results of such screening procedures as criteria in our study revolve around the fact that they are different from school district to school district and perhaps from school to school. For example, some districts rely on interviews by specialists, others use standardized testing. Still others arbitrarily place the child in a regular classroom and then ask the teacher to refer him or her to special programs as the need arises. Some districts focus only on English proficiency, others take into account proficiency in the home language as well. Of course, the labels attached to the results of the screening are also various. They include references to "English-language limitation", to "English independence", to "language dominance", etc.

Beyond the formal definitions of the screening procedures, there is the actual practice of them which can be of concern to a researcher. An external observer can only guess at the informal factors that might be operating to affect the screening processes. Are the bilingual services badly overcrowded? This could lead to lowering the implicit cutting point between LESA and non-LESA so as to provide

justification for placing more children in regular classrooms. Are the schools currently receiving funds on the basis of how many children need special services? That could lead to the opposite tendency -- screening procedures which would demand a high level of proficiency for a non-LESA classification. Does the faculty posit a dominant view of the mental capacities of a given ethnic group? And so on. It is virtually impossible to evaluate the extent to which such factors play a role in the way a given screening procedure is actually operated. What is clear, however, is that we can expect each school district to have its own unique screening procedure. Not only can we expect the cutting point between LESA and non-LESA to be variously placed in different school systems, but we can expect the continuum of English proficiency itself to be defined in various ways in the different locations. Thus, it would not be at all surprising to have the relationship between the screening procedures and our test and DORP be noticeably different in different locations.

What can be said about which school screening procedures is "better" than another? The research literature is not useful on this issue because there is no absolute scale or standard of English proficiency against which to compare them. The strategy adopted here was to ask various state education agencies to recommend local districts that had exemplary screening programs relative to our purposes. Then the local school districts were contacted directly and asked to participate in the study. Their participation was to consist of providing NCES with "the names and addresses of up to 500 children who have been screened, about half of whom have been determined to need special programs and half of whom have been determined not to need them" (from a letter to the superintendents of various school districts from NCES).

This method of obtaining samples differed markedly from the sampling methodology originally proposed by RTI and CAL in their proposals to NCES.

Those proposals suggested an informal cluster sampling procedure wherein interviewers would simply canvas neighborhoods known to contain high concentrations of the ethnic groups of interest. Screening questions would be asked upon first contact with a member of a household establishing the ethnic and linguistic backgrounds of the persons in the household. The interview would be continued, then, only for households that met certain screening conditions. During the course of interviews in the households of interest, permission would be sought to obtain information from the schools about the children in the household. After lengthy discussion during the week of June 16, it was decided that beginning with the schools and asking them to provide list samples was more efficient and more directly targeted on the objectives of the field test. It was also decided that NCES would make the contacts with the state and local education agencies.

Sampling in Different Age Ranges. Fisher's design specifications indicate that individuals of all ages were of interest to the Congress but that there was special interest in ages 5 to 17. However, NCES learned that screening programs and special curricula for secondary school students were largely non-existant or underdeveloped in most schools. The implicit philosophy seemed to be that helping the youngest children was most crucial and that older students either already knew a good deal of English or would learn it quickly given a minimum of assistance. As a result of this situation it was decided to limit the sampling of "children" to ages 5 - 13. This also coincided with the definition of "child" that was to be used in all other parts of the SIE questionnaire (i.e. the income and health-welfare sections); that is, in the SIE there were two questionnaires with some identical items, one to be asked of individuals 0 - 13 years and the other to be asked of individuals 14 and over. Thus, it would be particularly convenient to Census if the MELP could conform to that format as well. The letters sent to schools, then, asked for lists of children enrolled in elementary schools.

But what about the sampling of adults (14 years and older)? Was a known groups validation design possible for them? Did there already exist classifications of adults as being LESA and non-LESA? One source of such classifications might be adult education programs. Such programs routinely employ some sort of placement procedure for people with non-English backgrounds, and the resulting placement can be interpreted as a classification of an individual as either LESA or non-LESA. A difficulty with sampling from adult education programs is the self-selection factor. Clearly, those who voluntarily seek out an adult education program are not a random sample of any general population. Moreover, that population would not normally include any individuals between the ages of 14 and 18. Thus, adult education samples would exclude secondary-school students (who were also excluded from our child sample). Nevertheless, since no other a priori source of LESA and non-LESA categorizations could be found, the decision was made to ask school districts for "lists of names of up to 500 adults from foreign language backgrounds who are enrolled (or have been enrolled very recently) in adult basic education programs, including English as a second language if these are sponsored by your school district" (letter from NCES to school districts).

This, then, was the overall design of the field test as it evolved during the June discussions in San Francisco. The samples would be drawn from lists of prescreened children aged 5-13 provided to us by school districts with large concentrations of students having non-English language backgrounds. Separate lists of adult education program participants were also requested. The particular list from which an individual was drawn (LESA or non-LESA), then became a primary piece of criterion information about that individual along with his or her test score and DORP rating. (Interviewers were <u>not</u> informed of which list a respondent was on, i.e., all interviewing was done "blind" with respect to list membership.)

Choosing the Ethnic-Linguistic Groups to Participate in the Field Test. In RTI's proposal to NCES, field testing was suggested in the following groups: Cubans (Miami), Puerto Ricans (New York City), French (Manchester, New Hampshire), Chicanos (San Antonio), Navajos (Gallup, New Mexico), and Chinese (San Francisco). However, the revision of the sampling procedure required the reconsideration of all sites. Underlying the original choices was the requirement of sampling both from some of the largest groups in the U.S. having relatively high proportions of limited English speakers and from a culturally wide range of groups. Attempting to honor these requirements to as great an extent as possible, NCES approached the Texas, Florida, California, Arizona, New Mexico, New Jersey, and Massachusetts education agencies for their cooperation and suggestions about the school districts in their states would be most appropriate to approach for their cooperation. The Navajo Nation was also contacted for their suggestions. Negotiations for obtaining lists were begun with the Dade County (Miami), El Paso, Camden, San Francisco, Tuba City (Arizona), Window Rock (Arizona), and Ganado (Arizona) public school systems.

Eventually, lists of children were obtained from Dade County (Cubans), El Paso (Chicanos), San Francisco (Asians), Window Rock (Navajos) and Ganado (Navajos). The San Francisco Independent School District agreed to supply names of both Chinese and other Asian children in about equal numbers. Lists of adults enrolled in adult education programs were obtained only from Dade County and El Paso. Thus, the field test was held in four locations (Window Rock and Ganado are adjoining districts), and drew from five ethnic-linguistic groups -- Cubans, Chicanos, Navajos, Chinese, and other Asians.

A problem of finding adult respondents in the Navajo and Asian groups still remained. It was finally decided to sample adult respondents from the homes of the child respondents in those sites. This had the advantage of being cost-efficient

but had the disadvantage (from a sampling point of view) of only drawing adults from households containing children of elementary school age. The plan in those sites for selecting an adult respondent in a given household was as follows: first, the interviewer was to construct a household roster listing the name and age of each household member, and, second, she was to randomly choose one of the adults (age 14 and over) using a table of random numbers. This would give representation in the adult sample to all age groups over 13, including persons 14-18 who were not represented in the Cuban and Chicano samples.

2. The Accuracy of First-hand Data and "Proxy" Data

A focus of the field test was to investigate whether one adult in the household could give accurate answers to questions about another adult in the household, especially with regards to English proficiency. Such responses will be called proxy data and it was desirable to compare their quality, relative to the criterion measures, to the quality of first-hand data. This is important in the context of the SIE because of Census' preference for talking to only one adult in each household (the Household Respondent) and obtaining information about all members of the household from him. In order to address this question, interviewers were asked to obtain both first-hand and proxy responses to the MELP questionnaire whenever there were two adults present in the household.

3. The Language Ability of the Interviewer

Another concern about the accuracy of the data revolved around the fact that monolingual (English speaking) interviewers would inevitably be dealing with respondents whose English proficiency ranged from excellent to none. And, in addition to the linguistic factor there was also the cultural difference between the

monolingual, probably Anglo, interviewer and the ethnically distinct respondent. This difference could easily take its toll on the rapport between the two and thus influence the character of the data collected. In order to evaluate the severity of these problems, one component of the design of the field test was to compare the data collected by monolingual (English) interviewers and bilingual interviewers whose native language and ethnic origin was that of the respondent's. This was done by matching the assignments of monolingual with bilingual interviewers in each site through randomizing the names and addresses of the individuals they were to interview.

4. The Interviewing Procedures

All data collection and analysis activities associated with the field test, from the recruiting of interviewers to the statistical analysis of the data were the responsibility of the Research Triangle Institute under a subcontract arrangement with CAL. The following description of the field procedures is taken from pages 24-27 of RTI's final report of their subcontract activities. "The "QQ" referred to is the Census-style questionnaire containing verious demographic and candidate MELP questions.

> Interviewer assignments were prepared by the site supervisory teams, following detailed procedures designed by RTI's Sampling Department to (1) equalize the effort for children and adults; (2) equalize the effort for each child or adult's proficiency level defined by the schools (e.g., in Miami: non-independent, intermediate, and independent); (3) increase the precision of the comparison between bilingual and monolingual interviewers; and (4) randomize the subsample of interviews to be monitored by the CAL staff.
>
> The field procedures followed by the interviewers during the field test are detailed in the interviewer's field manual, a copy of which is included in the attachment to this report. The procedures for the three principal types of cases are summarized below:
>
> . Designated Child Respondents (DCRs)

(1) The interviewer calls in person at the sample household at a time when a household respondent (household member at least 14 years old) is likely to be home.

(2) The interviewer locates a household respondent and (a) introduces herself, (b) verifies that the DCR is a household member, and (c) explains the study.

(3) The interviewer administers the Census Questionnaire (CQ) and Household Information Form (HIF) to the Household Respondent. (NOTE: The household respondent responds to the CQ on behalf of the DCR.)

(4) The interviewer determines the age of the DCR.

(5) The interviewer interviews the DCR. (NOTE: If the DCR is ten or older, the interviewer administers the CQ and criterion measures; if the DCR is nine or younger, the interviewer administers only the criterion measures.)

. Designated Adult Respondents (DARs) from School Lists (Miami and El Paso)

(1) The interviewer locates a household respondent as for DCRs above.

(NOTE: The household respondent can also be the DAR, if the DAR is the first person 14 or older the interviewer encounters.)

(2) The interviewer administers the CQ and HIF to the household respondent.

(NOTE: The CQ is second-hand if the household respondent is not also the DAR; first-hand if the household respondent is the DAR.)

(3) The interviewer interviews the DAR.

(NOTE: If the household respondent is the DAR, the CQ will have already been administered and the interviewer continues with the criterion measures.)

. Designated Adult Respondents (DARs) Randomly Selected from DCR Households (N.E. Arizona and San Francisco)

The interviewer locates a household respondent, as above. The interviewer then randomly selects an adult member of the household, who becomes the DAR. The interviewer then proceeds to interview the household respondent, DCR, and DAR as described above.

A number of minor procedural changes and refinements were made as the fieldwork progressed and problems became apparent. One notable change that was implemented near the end of the fieldwork period concerned obtaining

second-hand CQ information on adults. In order to increase the number of cases where second-hand CQ data were obtained on DARs, interviewers were instructed to attempt to find a household respondent who was not also a DAR. One callback was authorized to accomplish this, if necessary.

Respondents were paid cash incentives by the interviewers at the rate of $2.00 for each completed CQ and $2.00 for each completed set of criterion measures. Incentive payments made directly to DCRs were made with the knowledge of a responsible adult member of the household. No payment was made for the short HIF, which was completed in conjunction with the initial CQ.

Interviewers were instructed to make up to two calls at a sample household in order to contact a household respondent. If the interviewer was unable to contact a household respondent on the first call, she would attempt to find out from neighbors when the household residents were most likely to be found at home, and made her second call at that time. If neighbor information was unavailable, the interviewers were instructed to make the return call after 6:00 p.m. on a weekday or on a weekend. After initial contact, the interviewer was allowed up to two or more calls to complete interviewing in the household. If she had still not completed her work at the household after two additional callbacks, she was instructed to discuss the case with a site supervisor immediately.

The interviewers were not permitted to substitute non-sample persons for designated respondents. All non-interview cases had to be discussed with a site supervisor, who would determine what, if any, additional action should be taken. If no further action was warranted, the supervisor would approve the noninterview result and provide the interviewer with a substitute case, according to the interviewer assignment procedures developed by RTI's Sampling Department.

The two RTI supervisors in each site remained in the field during the fieldwork period in order to monitor closely the data collection activities of the interviewers. The supervisors normally met with each interviewer at least twice a week to review the status of each of her active cases and to advise and assist her as necessary. The supervisors were responsible for editing and approving the instruments associated with each completed case and for mailing completed cases to RTI on a flow basis. Additional cases were assigned to interviewers when appropriate, following procedures specified by RTI's Sampling Department. The supervisors were also responsible for validating the fieldwork by contacting at least ten percent of each interviewer's respondents (those not monitored by CAL staff) to verify that the interviewer had conducted the interview properly and that the respondents had been paid. Other responsibilities of the site supervisors included monitoring interviewer costs; controlling the issuing and retrieving of advances to interviewers for use in making cash payments to respondents; recruiting and training replacement interviewers, as necessary; maintaining records on the handling and status of each case; and reporting to RTI at least weekly the status of the fieldwork in the field test site.

Interviewer training and interviewing were begun in Miami and El Paso on July 21 and on July 28 in Arizona and San Francisco. Data collection was completed on August 16 in Miami and El Paso and on August 23 in Arizona and San Francisco. The results of these efforts are discussed in detail in Section VI.H of RTI's final report, but Table 1, reproduced from that report, summarizes statistics on the numbers of interviews attempted and completed in each site, along with measures of the amount of effort expended to obtain them.

(Refer to Table 1, on next page)

5. Monitoring of Interviews

CAL personnel monitored approximately 15% of the interviews in each site for two reasons:

1. To observe and report on the interaction between interviewer and respondent during the asking and answering of each potential MELP question for evaluating and improving the questions.
2. To ensure that the interviewers were following recommended procedures and, if necessary to recommend any modifications of those procedures to RTI and CAL supervisory personnel.

CAL monitors were randomly assigned to interviewers on a daily basis and simply accompanied the interviewer on his or her rounds for the day. The behavior observation system described in Chapter III was filled out for each administration of the "CQ" -- first hand or proxy. Upon the completion of the field work, each monitor submitted a summary report, either written or verbal, focused on the aspects of the interview procedure that seemed to work well, those that worked badly, etc.

Table 1:

DATA COLLECTION RESULTS OF FIELD TEST[1/]

	Miami	El Paso	Arizona	San Francisco	Total
Potential Respondents Assigned[2/]	1,079	1,071	972	1,192	4,314
Interviews with Children	335	426	358	353	1,472
Interviews with Adults	333	265	315	319	1,232
Total Interviews (Percent)	668 (62%)	691 (65%)	673 (69%)	672 (56%)	2,704 (63%)
Refused (Percent)	26 (2%)	18 (2%)	16 (2%)	54 (5%)	114 (3%)
Other Nonrespondents[3/] (Percent)	385 (36%)	362 (34%)	283 (29%)	471 (40%)	1,501 (35%)
Total Nonrespondents (Percent)	411 (38%)	380 (35%)	299 (31%)	525 (44%)	1,615 (37%)
Total Hours Charged[4/]	2,916	2,992	3,203	2,917	12,028
Total Miles Driven[5/]	22,966	21,079	34,328	8,299	86,672
Average Hours Per Interview	4.4	4.3	4.8	4.3	4.5
Average Miles Per Interview	34.4	30.5	51.0	12.4	32.1
% of Adult Respondents with 2nd Hand Census Questionnaires[6/]	36%	36%	83%	36%	48%

1/ Figures in this table are based upon manual counts and computations by interviewers and supervisors and have not been verified by machine tabulations.

2/ In Miami and El Paso both children and adults were assigned to interviewers. In Arizona and San Francisco only children were assigned, since no adult lists were obtained for these sites. Interviewers randomly selected an adult from each sample child's household in these sites. For Arizona and San Francisco, therefore, the number of potential respondents was twice the number of sample children assigned.

3/ Examples of "other" nonrespondents include cases where the sample member had moved to another city; where the address was nonexistent; where the sample member could not be contacted at home in the prescribed number of interviewer visits; where the sample member was out of town; or where he was sick, institutionalized, or otherwise unavailable.

4/ Includes training time.

5/ Includes mileage incurred in connection with training.

6/ Figures shown indicate the percent of adult respondents in each site about whom Census Questionnaire data were obtained from a household member other than the respondent as well as from the respondent himself.

The following CAL staff were assigned to the various sites:

Dade County

Dr. Alberto Rey - CAL site supervisor
Pedro Ruiz
Cynthia Lindsey
Roberta Mailman

El Paso

Amador Bustos - CAL site supervisor
Dr. J. Terry Webb
Gloria Lozano
Benjamin Zambalas

Arizona

Carolyn Karelitz - CAL site supervisor
Evangeline Kamitsuka
Annie Panlibuton
Claire McKenzie

San Francisco

Anna Lai
Michael SamVargas > CAL site supervisors
Jennie Yee
Margaret Robbins

BEST COPY AVAILABLE

6. Visits to the Sites by CAL Central Staff

During the course of the field work, each site was visited by at least one member of the CAL central staff. The objects of these trips were:

1. To interview CAL and RTI field personnel in depth to learn about and resolve any procedural or coordination difficulties in the two staffs.
2. To interview local school officials in depth to gather information relevant to the screening procedures which formed the basis for the list samples.

The trips made were:

Miami: Robert Pearl (CAL consultant)
Jeanne Freeman
Walter Stolz

El Paso: Jeanne Freeman

Arizona: Walter Stolz

San Francisco: Rudolph Troike

7. Editing, Coding, and Entering the Data into Computerized Files.

The details of this process may be found in Section IV.I of RTI's final report. Basically, the procedure involved several stages of checking and editing the completed interview materials and then entering the data directly into computerized files through the use of a terminal. The confidentiality procedures employed during these phases of the work are described in Section IV.J of RTI's report. The data entry procedures were completed during the week of September 8. All of the statistical analyses performed on these data were implemented by the RTI statistical staff under CAL's direction.

V. Preliminary Analyses: Selection of the MELP Questions

From the point of view of Census Bureau field operations, the optimal MELP was a small set of simple questions which could be asked by the interviewer about each member of the household. Ideally, all such information would be obtained from the Household Respondent. As conversations with Census and NCES progressed during the first several months of the project, it became very clear that any direct measure of proficiency, such as an interviewer-administered rating, which required the interviewer to actually talk with each person for whom a LESA or non-LESA categorization was to be made, would require extensive replanning and re-budgeting on the part of Census. Thus, the obvious first priority of the analysis of the field test data was to ascertain the degree of relationship between individual MELP questions and the criterion variables. If several of them showed relatively high and consistent relationships with the criteria across all groups, then some "mapping" of those questions onto LESA and non-LESA categories was clearly the MELP of choice. This chapter summarizes the relationships of the various individual MELP questions to the criteria. In fact, high and stable (across groups) relationships were found and thus a set of such questions was forwarded to NCES on October 2, 1975 for use in the SIE. Also covered in this chapter are the rules used to quantify the responses to the MELP questions for further statistical analysis.

The remainder of the project work, then, was devoted to constructing "scoring keys" for these questions -- that is, procedures for categorizing an individual as LESA or non-LESA on the basis of his quantified responses to the MELP questions. Those activities and their results are summarized in Chapters VII and VIII.

1. "Cleaning" the Data Files

Before any analyses of the field test data were done, the files were examined so that any data gathered from respondents who were irrelevant to the project could

be eliminated. In particular, only the data from people with non-English language backgrounds were appropriate to be analyzed since only they would be administered the MELP in the SIE. Therefore, the data from respondents who met all three of the following conditions were eliminated permanently from the data files:

a. No other language but English present in the household.

b. The respondent spoke no other language but English.

c. The respondent was born in the U.S.

The data from 40 children and 14 adults were eliminated from the study as a result of this procedure.*

2. Relationships of Individual Questions to the Criteria

All analyses were accomplished within the framework of the SPSS statistical system. The basic analysis device was a simple contingency table where the responses to each census question were cross-tabulated with test total scores and list information (where available) separately for each of the populations represented in the field test as follows:

a. Children:

1) Cubans

2) Chicanos

3) Chinese

4) Other Asians

5) Navajos from Ganado schools

6) Navajos from Window Rock schools

b. Adults:

1) Cubans

2) Chicanos

* It was later ascertained that most of the children who were eliminated were from monolingual families who had requested placement in the bilingual program to learn the non-English language.

3) Chinese

4) Other Asians

5) Navajos

The Navajo children were split by school district when their data were cross-tabulated with school list because the two school districts from which the field test sample was drawn had very different methods of assigning children to lists. School list information was only available for Cubans and Chicanos.

All contingency tables that included test scores were constructed by arbitrarily dividing the test scores into ten-point intervals. The possible range for the children's test was 0-67, the possible range for the adult's test was 0-57.

For each two-way cross tabulation (question responses by list or test for a given subpopulation), several summary statistics were computed. On the recommendation of Dr. Robert Mason of RTI, the two indices used were Cramer's V (Cramer, 1945) and the correlation ratio, eta. The former was used where the responses to a question were not orderable on a continuum (e.g., origin or descent), while eta was used when the response categories were ordered. In the latter case the eta was computed using the question responses as predictors and test or list as the predicted variable.

To facilitate the examination of the several hundred cross-tabulations, a two day conference was convened of the following individuals:

Burton Fisher, University of Wisconsin

John Upshur, University of Michigan

Protase Woodford, Educational Testing Service

Harold Yee, Asian Inc. (San Francisco)

Robert Mason, RTI

Alberto Rey, Howard University and CAL

Margaret Bruck, McGill University and CAL

G. Richard Tucker, McGill University

Walter Stolz, CAL

Leslie Silverman, NCES

Vicki Kojsich, NCES

David Orr, NCES

Included in this group were specialists in language testing, survey research, statistics, linguistics, psychometrics and bilingualism. In addition, three of the specialists were members of three of the largest ethnic groups to be surveyed by the SIE.

The conference was held September 22-24 at CAL, with Leann Parker and Evangeline Kamitsuka providing logistical support. Although the discussion of the data ranged over many topics during the two days, the basic question selection procedure used by the group was as follows:

1. Summary tables were created (separately for children and adults) in which only the Cramer's V and/or the eta was entered for each question/criterion-measure/subpopulation combination.

2. Questions with consistently high indices of association were selected for further examination. Generally speaking, for a question to be selected, its Cramer's V values had to exceed .20 in every subpopulation (except Window Rock when the cross-tabulation was with list).

3. The cross tabulations for the selected questions were examined to make sure that the pattern of association between the question responses and criterion was the same within all subpopulations.

4. The data for the discarded questions were perused once more to ascertain that the question had not been wrongly eliminated.

The summary tables from which the group worked are reproduced as tables 1 through 4.* Underlined rows correspond to questions recommended to NCES as MELP questions on October 2, 1975. The field test questionnaire is reprinted as Figure 1 and the final wordings of the MELP questions as recommended to NCES are given in Section 4 of this chapter.

Comments on Tables 1 and 2:

1. It was assumed that questions 1, 2, 3, 4 and 21 would be present in the SIE questionnaire regardless of their usefulness as LESA indicators and thus they were not included in the recommended MELP questions even though most of them were highly related to the criteria.

2. Question 5 was retained as one of the MELP items proposed for inclusion because it was part of questions 6 and 7. (Its relationships to the criteria were low because virtually no children were characterized by the household respondents as neither speaking nor understanding <u>any</u> English.)

3. Question 27 was another way of phrasing questions 5, 6, and 7. It had been used in the NCES supplement to the July CPS and so was used here, but it was judged more difficult to understand than 5, 6, and 7 and so was not selected for the final MELP.

4. For Cubans, the relationship of question 31 to the criteria was low because the household language was almost universally Spanish in that group.

With respect to tables 3 and 4 it should be noted that the relationships between the questions and the adult's list classification are generally lower than between the questions and test scores.

---

* Relationships of questions to DORP scores were also inspected by the group during the selection process, but because of incomplete data they did not play a central role in the selection.

Table 1: Children: Cross Tabulations of Responses to Questions with Test Total Scores

Numbers are Cramer's V, except where * appears after questions, Etas are given for asterisked questions.

MELP Questions	Cubans	Chicanos	Asian (Exc. Chin.)	Chinese	Navajos
1.	415	147	419	654	416
2.	202	200	185	133	141
3.	217	249	119	164	113
4. *	447	567	274	471	---
5. *	247	237	092	140	256
6. *	625	636	523	544	509
7. *	634	616	518	491	402
9.	128	176	285	133	159
10. *	163	327	286	272	368
11. *	150	351	346	120	340
12.					
a.	256	286	219	216	535
b.	197	380	179	249	624
c.	147	318	202	199	305
d.	278	326	253	247	289
13. *	263	385	239	238	234
14. *					
a.	246	234	315	353	308
b.	410	262	249	322	287
c.	469	347	340	497	259
15. *	118	208	349	060	088
16. *	211	117	000	156	020
17. *	103	050	092	045	064
18. *	107	011	183	174	119

Table 1 continued.

MELP Questions	Cubans	Chicanos	Asian(Exc. Chin)	Chinese	Navajos
19. *	163	295	413	070	250
20. *	033	195	079	049	092
21. *	526	209	399	650	538
22. *	602	246	458	584	474
23. *	119	512	180	235	437
24. *	174	518	135	395	380
25. *	163	034	036	032	085
27.	281	310	296	262	274
28. *	345	302	146	289	235
31.	128	469	246	208	320
32.	167	406	208	213	342

Table 2 - Children: Crosstabulations of Responses to Questions with School List Information

Numbers are Cramer's V except where * appears after question; Etas are given for asterisked questions.

MELP Questions	Cubans	Chicanos	Asian(Exc. Chin)	Chinese	(Gan.) Navajo	(WR) Navajo
1.	226	031	011	240	173	048
2.	154	403	267	---	223	089
3.	140	337	429	248	123	034
4. *	561	616	377	277	---	---
5. *	172	150	103	148	070	087
6. *	580	659	347	537	420	133
7. *	516	657	378	498	360	257
9.	076	282	416	241	148	158
10. *	100	522	354	399	420	234
11. *	160	479	306	312	353	392
12. a.	250	698	384	379	243	267
b.	076	750	403	422	263	270
c.	117	607	368	234	228	206
d.	258	482	214	360	269	235
13. *	268	418	322	274	326	142
14. a. *	118	154	159	486	123	114
b. *	257	165	234	355	213	047
c. *	281	191	391	301	189	066
15. *	115	258	227	197	128	096
16. *	146	075	103	120	120	010
17. *	053	034	117	031	149	127
18. *	029	044	162	366	041	067
19. *	165	247	353	271	362	109

Table 2 continued.

MELP Questions	Cubans	Chicanos	Asian(Exc. Chin)	Chinese	(Can.) Navajo	(WR) Navajo
20. *	075	163	128	161	125	191
21. *	307	126	263	326	263	495
22. *	531	121	360	376	187	484
23. *	120	626	334	283	299	124
24. *	309	665	328	487	292	264
25. *	056	094	159	044	190	116
26.	219	208	764	378	385	456
27.	279	536	295	368	359	105
28. *	216	309	093	245	161	173
29. *	045	048	040	---	136	080
31.	068	717	426	376	312	123
32.	154	667	423	373	242	117

Table 3 - Adults: Crosstabulations of Responses to Questions with Test Total Scores

Numbers are Cramer's V except where * appears after question; Etas are given for asterisked questions.

MELP Questions	Cubans	Chicanos	Asian (Exc. Chin.)	Chinese	Navajo
1.	166	051	392	316	212
2.	104	183	280	213	168
3.	135	153	331	298	142
4. *	225	235	279	270	000
5. *	376	311	105	547	288
6. *	561	477	534	703	645
7. *	519	467	565	672	592
9.	110	115	371	243	102
10. *	150	147	496	180	220
11. *	157	165	456	333	224
12.					
a.	120	426	373	351	253
b.	162	135	347	324	191
c.	159	214	322	308	143
d.	201	198	360	386	215
e.	208	145	286	338	269
13. *	281	347	336	361	425
14.					
a. *	450	295	389	578	564
b. *	493	388	454	562	382
c. *	399	266	366	620	434
15. *	113	116	213	428	110
16. *	069	175	130	424	143
17. *	154	039	016	051	145

Table 3 continued.

MELP Questions	Cubans	Chicanos	Asian (Exc. Chin.)	Chinese	Navajo
18. *	133	141	074	086	253
19. *	113	262	279	262	120
20. *	091	057	168	051	041
21. *	474	348	512	666	715
22. *	365	412	581	668	667
23. *	143	190	276	306	263
24. *	205	320	456	616	287
25. *	009	157	106	051	051
26.	691	438	707	829	543
27.	290	180	366	416	314
28. *	240	200	347	253	389
29. *	161	094	197	103	251
30.-E	191	301	258	298	321
31.	106	185	271	360	407
32.	219	177	425	301	284

Table 4: Adults: Crosstabulations of Responses to Questions with School List Information.

Numbers are Cramer's V except where * appears after question; Etas are given for asterisked. questions.

MELP Questions	Cubans	Chicanos
1.	151	023
2.	133	198
3.	093	109
4. *	064	177
5. *	255	058
6. *	416	229
7. *	321	113
9.	089	029
10. *	125	129
11. *	150	102
12.		
a.	082	107
b.	078	127
c.	082	152
d.	083	143
e.	085	129
13. *	183	106
14. *		
a.	334	161
b.	350	144
c.	331	138
15. *	058	075
16. *	105	092

Table 4 continued.

MELP Questions	Cubans	Chicanos
17. *	131	067
18. *	097	044
19. *	100	107
20. *	148	159
21. *	148	159
22. *	318	384
23. *	138	100
24. *	161	247
25. *	054	085
26.	671 †	406 †
27.	237	137
28. *	138	070
29. *	255	074
30-E.	023	040
31.	098	002
32.	---	076

† Based on very small sample sizes

3. An Evaluation of the MELP Questions: Reports from the Monitors

Once the questions had been selected they were examined to see if they needed to be improved in their wordings. One source of information relevant to this was the monitors' observation data and their summary reports submitted at the end of the field test. The table below gives the results of the monitor observation system for several of the questions.

Behavioral Category	Question number									
	2	5	6	7	12b	12d	21	22	27	31
No Response*	0	0	0	0	0	0	0	0	0	0
Irrelevant Answer*	0	0	0	0	0	0	1	0	1	0
Another person Answers*	7	6	9	8	6	9	11	9	8	7
Seeks clarification*	14	1	4	3	2	0	9	6	7	3
Interv. Rephrases*	20	2	7	5	3	5	14	14	14	7
Interv. uses Native L.*	36	35	30	30	33	34	35	36	37	33
Respondent uses Native L.*	36	35	31	31	35	34	36	35	37	33
Total Frequencies	376	371	334	333	362	361	366	349	360	366
Sum of N.R., I.A., S.C.,I.R.*	34	3	11	8	5	5	24	20	22	10

* Percents of total frequencies.

These are pooled accross all administrations of the MELP questions that were monitored. The last row gives the total percent of no responses, irrelevant answers, seeks clarification, and interviewer rephrases and might be taken as a general index of the difficulty of administration of the question. The troublesome questions were clearly: #2 (origin and descent), #21 (level of education), #22 (years of education in English), and #27 (CPS question rating English proficiency). Comments from the monitors indicated that:

1. For question 2, the words "origin" and "descent" as well as the concept of ethnic background often caused difficulty. Navajos needed to have the word "tribe" substituted.
2. In question 21, there was often uncertainty about how to translate foreign schooling into U.S. terms.
3. In question 22, there was sometimes an ambiguity between years of having been taught the English language and years of instruction in content areas <u>using</u> English as the medium of instruction. The latter was intended.
4. Question 27 was double barreled and the alternative responses were extremely difficult to understand.
5. In responding to Question 31, some respondents indicated that both languages were used equally often and they had to be prodded into making a forced choice.
6. For questions 6 and 7 most problems involved the term "adequately".
7. Finally, it was suggested that question 7 be placed before question 6 because often the word "speak" was initially taken in its generic sense meaning both speak and understand. However, if the question about understanding was placed first, the proper sense of "speak" would be suggested to most respondents.

Originally we anticipated that the two categories "interviewer uses native language" and "respondent uses native language" would be indicative of difficulties in communicating a question or an answer in English. This may have been the case for the monolingual English-speaking interviewers, but according to the monitors' comments, it was not the case in the interviews conducted by bilingual interviewers. In the latter case, the interviewers found that it was viewed by respondents as a lack of courtesy for the interviewer to attempt to conduct the interview in English (as was their instruction) when it was difficult and/or embarrassing for the respondent to do so and when the interviewer was clearly competent in the respondent's native language. Thus, interviews were frequently conducted in the native language even when, according to the monitor's judgment, it could have been conducted mostly or entirely in English. Accordingly, these behavioral categories were not interpreted as originally planned.

4. Modifications to the "How Well" Questions

From the beginning of the field test it was clear that the set of response alternatives to the "how well" questions (#6 and #7) could be improved. After a week of field testing with the set very well, well, adequately, just a little, and not at all, the term adequately was replaced by two alternatives: adequately for most purposes and adequately for only a few purposes. (CAL staff considered adequately to be overly ambiguous.) However, this did not solve the problem. The word remained highly ambiguous to some, and to others it was simply unfamiliar.

Also, the term well proved to be non-discriminative. In fact, analysis of the data showed that well, adequately for most purposes, and adequately were all applied to people of about the same English proficiency level as measured by test score. Table 5 gives the mean test score for respondents to whom each response alternative was applied. For example, the mean test score of all adults who rated

themselves as speaking English "very well" was 43, and the mean test score of all children who were rated as speaking "very well" was 56.

Table 5: Mean test scores for each response alternative in the question rating English proficiency. (Pooled across ethnic groups)

Response Alternative	Adult		Child	
	Speak	Understand	Speak	Understand
Very Well	43	41	56	55
Well	34	33	49	52
Adequately for most	34	31	48	47
Adequately	32	29	46	46
Adequately for few	28	25	40	39
Just a little	17	17	37	36

For adults, the average difference between the means of well, adequately for most purposes, and adequately was 1.5 compared with an average difference of 7.3 between all other successive alternatives. The largest difference between any successive pair of the three was 2.25 while the average difference between all other successive pairs was 4.83. On the basis of this analysis, it was decided to collapse the three alternatives into a single scale position. After consultation with a number of the CAL staff, the following response alternatives were agreed upon and included in CAL's October 2 memorandum to NCES:

1. Very well
2. All right
3. Enough to get by
4. Just a few words
5. Not at all

FIGURE 1

(items selected for MELP are starred)

BILINGUAL STUDY

CENSUS QUESTIONNAIRE

O.M.B. No. _____

Expires _____

ID No. of DR _____ Sex _____

FI _____ FI No. _____ Date ____

Type (✓): ☐ Self Report ☐ Second Hand Report

* 1. What is . . .'s date of birth?

Month _____ Day _____ Year _____

2. What is . . .'s origin or descent? *(USE FLASH CARD A)*

_____

* 3. In what state or foreign country was . . . born? *(USE FLASH CARD B)*

* 4. When did . . . come to the U. S. to stay?

_____ 1. 1975
_____ 2. 1973-1974
_____ 3. 1971-1972
_____ 4. 1966-1970
_____ 5. 1961-1965
_____ 6. Before 1961
_____ 7. Don't know

* 5. Does . . . speak or understand <u>any</u> English?

_____ 1. Yes
_____ 2. No *(SKIP TO Q.8)*
_____ 3. Don't know *(SKIP TO Q.8)*

* 6. How well does . . . speak English? *(READ ANSWER CHOICES 1-5)*

_____ 1. Very well
_____ 2. Well
_____ 3. Adequately
_____ 4. Just a little
_____ 5. Not at all
_____ 6. Don't know

* 7. How well does . . . understand spoken English? *(READ ANSWER CHOICES 1-5)*

_____ 1. Very well
_____ 2. Well
_____ 3. Adequately
_____ 4. Just a little
_____ 5. Not at all
_____ 6. Don't know

8. What *(OTHER)* languages does . . . speak? *(USE FLASH CARD C)*

_____

(IF NONE, SKIP TO Q.12. IF ONLY ONE, SKIP TO Q.10)

9. Which of these languages does . . . speak most often? *(USE FLASH CARD C)*

_____

10. How well does . . . speak *(PRINCIPAL LANGUAGE FROM Q.8 OR Q.9)?*
(READ ANSWER CHOICES 1-4)

_____ 1. Very well
_____ 2. Well
_____ 3. Adequately
_____ 4. Just a little
_____ 5. Don't know

11. How well does . . . understand *(PRINCIPAL LANGUAGE FROM Q.8 OR Q.9)?*
(READ ANSWER CHOICES 1-4)

_____ 1. Very well
_____ 2. Well
_____ 3. Adequately
_____ 4. Just a little
_____ 5. Don't know

12. What language does . . . usually speak when talking to: *(USE FLASH CARD C)*

* a. brothers and sisters? _____
b. parents? _____
c. other older relatives? _____
* d. . . .'s best friend? _____
* e. *(IF . . . IS AN ADULT)* children in the household? _____

13. During the past year, did . . . have difficulty reading books because they were in English?

_____ 1. Yes
_____ 2. No
_____ 3. Don't know

14. How often does . . . read:

* a. an English-language newspaper? *(READ ANSWER CHOICES)*

_____ 1. Often
_____ 2. Occasionally
_____ 3. Not at all

b. magazines in English? *(READ ANSWER CHOICES)*

_____ 1. Often
_____ 2. Occasionally
_____ 3. Not at all

c. books in English? *(READ ANSWER CHOICES)*

_____ 1. Often
_____ 2. Occasionally
_____ 3. Not at all

15. How often does . . . read newspapers, magazines, or books in a language other than English? *(READ ANSWER CHOICES)*

_____ 1. Often
_____ 2. Occasionally
_____ 3. Not at all

16. At any time during the past year, did . . . attend regular school in the U. S.?

_____ 1. Yes
_____ 2. No
_____ 3. Don't know

17. During the past year, did . . . take any courses at business, vocational or technical school?

_____ 1. Yes
_____ 2. No
_____ 3. Don't know

(IF "NO" OR "DON"T KNOW" TO BOTH Q's 16 AND 17, SKIP TO Q.20)

18. In any school or course attended during the past year, was . . . taught in a language other than English?

_____ 1. Yes
_____ 2. No
_____ 3. Don't know

19. During the past year has a teacher, counselor, or school official said that . . . had difficulty speaking or understanding English?

_____ 1. Yes
_____ 2. No
_____ 3. Don't know

20. At any time during the past year did . . . take any course or class for people whose principal language is not English?

_____ 1. Yes
_____ 2. No
_____ 3. Don't know

* 21. What is the highest grade or year of regular school . . . has ever attended? *(USE FLASH CARD D)*

_____

(IF "NONE" SKIP TO 27. IF "DON'T KNOW," SKIP TO Q.23)

* 22. How many years of . . .'s schooling was taught in English? _____

23. Did . . . speak English before going to school for the very first time?

_____ 1. Yes
_____ 2. No *(SKIP TO Q.25)*
_____ 3. Don't know *(SKIP TO Q.25)*

24. How well did . . . speak English before going to school for the very first time? *(READ ANSWER CHOICES 1-4)*

_____ 1. Very well
_____ 2. Well
_____ 3. Adequately
_____ 4. Just a little
_____ 5. Don't know

25. Has . . . ever repeated a grade in school?

_____ 1. Yes
_____ 2. No *(SKIP TO Q.27)*
_____ 3. Don't know *(SKIP TO Q.27)*

26. What grade(s) did . . . repeat? _____

27. Does . . . have any difficulty in speaking or understanding English? *(READ ANSWER CHOICES)*

_____ 1. Yes, difficulty in both speaking and understanding
_____ 2. Yes, difficulty only in speaking
_____ 3. Yes, difficulty only in understanding
_____ 4. Yes, doesn't speak or understand at all
_____ 5. No, no difficulty in speaking or understanding
_____ 6. Don't know

28. Does . . . prefer to avoid places where only English is spoken?

_____ 1. Yes
_____ 2. No
_____ 3. Don't know

29. During the past year has . . . been employed at any time?

_____ 1. Yes
_____ 2. No *(SKIP TO Q.31)*
_____ 3. Don't know *(SKIP TO Q.31)*

30A. For whom did . . . work? *(NAME OF COMPANY, BUSINESS, ORGANIZATION, OR OTHER EMPLOYER)*

_____

30B. What kind of business or industry is this? *(FOR EXAMPLE, TV AND RADIO MANUFACTURING, RETAIL SHOE STORE, STATE LABOR DEPARTMENT, FARM)*

_____

30C. What kind of work did . . . do? *(FOR EXAMPLE, ELECTRICAL ENGINEER, STOCK CLERK, TYPIST, FARMER.)*

_____

30D. What were . . .'s most important activities or duties? *(FOR EXAMPLE, TYPES, KEEPS ACCOUNT BOOKS, FILES, SELLS CARS, OPERATES PRINTING PRESS, FINISHES CONCRETE)*

_____

30E. At work, what language does . . . usually speak? *(USE FLASH CARD C)*

_____

* 31. What is the usual language spoken in this household? *(USE FLASH CARD C)*

_____

32. What other languages are spoken in this household? *(USE FLASH CARD C)*.

_____

The question of whether to have a separate screening item such as "Does . . . . speak or understand any English?" or to have the not at all alternative of the "how well" items characterize them was left to NCES. It was found that few adults or children (10% and 2% respectively) were recorded as neither speaking nor understanding any English, and thus justification as to whether question 5 should be retained was left to the designers of the final SIE questionnaire. Such a question could be useful more as a device for moving to a new topic than for the information it yields by itself.

5. MELP Questions as Recommended to NCES

On October 2, 1975 the following questions were recommended to NCES for inclusion in the SIE instrument.

A. "How well" questions:

1. Does . . . speak or understand <u>any</u> English?

____ 1. Yes

____ 2. No

____ 3. Don't know

2. How well does . . . understand spoken English?

____ 1. Very well

____ 2. All right

____ 3. Enough to get by

____ 4. Just a few words

____ 5. Not at all

3. How well does . . . speak English?

____ 1. Very well

____ 2. All right

____ 3. Enough to get by

____ 4. Just a few words

____ 5. Not at all

B. English usage questions:

1. What is the usual language spoken in this household? (To be asked only once of the household respondent; interviewer coded for each member of the household.)

2. What language does . . . usually speak when talking to:

   a. brothers and sisters? (children only)

   b. . . .'s best friend?

C. Questions about reading habits:

1. How often does . . . read an English-language newspaper? (Adults only)

   ____ 1. Often

   ____ 2. Occasionally

   ____ 3. Not at all

D. Educational questions

1. How many years of . . .'s schooling was taught in English?

II. Questions forwarded for inclusion in the SIE questionnaire on the recommendation of the Language Group Representatives.

1. How well does . . . understand spoken [principal non-English language (from III, 8a and b)]?

   ____ 1. Very well

   ____ 2. All right

   ____ 3. Enough to get by

   ____ 4. Just a few words

   ____ 5. Not at all

2. How well does . . . speak [principal non-English language]?

____ 1. Very well

____ 2. All right

____ 3. Enough to get by

____ 4. Just a few words

____ 5. Not at all

III Non-MELP questions: It was our understanding that the following question would be asked for reasons other than to categorize individuals as LESA or not: however we assumed that they would be available for incorporation into the MELP.

1. What is . . .'s date of birth

2. What is . . .'s origin or descent ("tribe" if Native American)?

3. In what state, U.S. territory, or foreign country was . . . born?

4. When did . . . come to the U.S. mainland to stay? [Skip if answer to preceding question was "this state" or "different state".]

5. How many years of . . .'s schooling was not on the U.S. mainland?

6. What is the highest grade or year of regular school . . . has ever attended?

7. What other languages are spoken in this household? (to follow question B1)

8. a. What other languages (besides English) does . . . speak?

   b. Which of these languages does . . . speak most often?

6. Definitions of the MELP Variables

Once the MELP questions had been selected for use in the SIE, there remained the task of quantifying the responses to them so that they could be entered into further statistical analyses to derive one or more "scoring keys". Such scoring keys would determine how any given individual would be actually classified as LESA or not on the basis of his MELP responses. The quantified responses to the MELP questions will be called the MELP variables. There were ten MELP variables for children and 11 for adults. They are defined below. The labels in capital letters will be used to refer to the various MELP variables henceforth. Questionnaire numbers refer to those in Figure 1.

Child MELP Variables

A. Length of time in U.S. (WHEN): This variable was a composite of questionnaire items #3 and #4, and it had three possible values.

1 - Born outside the U.S. and came to U.S. after 1972
2 - Born outside the U.S. and came to U.S. before 1973
3 - Born in the U.S

B. Rating of proficiency in Speaking English (SPEAK): Derived from items #4 and #5, and scored on a scale of 1 through 5:

1 - Does not speak any English at all
2 - Speaks just a little
3 - Speaks adequately for a few purposes
4 - Speaks adequately; adequately for most purposes, or well
5 - Speaks very well

Any missing data were given the value of 2.

C. Rating of proficiency in understanding spoken English (UNDERSTAND): Also scored on a 1 to 5 scale using the same scale labels as SPEAK

only with the word "understand" replacing each occurrence of "speak." Derived from items # 4 and 6. Any missing data were given the value of 2.

D. Usual language spoken in the household (HLANG): This was a three-valued variable derived from item #31.

1 - not English
2 - any missing data
3 - English

E. Usual language spoken with brothers and sisters (SIB): Scored exactly as was HLANG. Derived from item # 12a.

F. Usual language spoken with best friend (FRIEND): Scored exactly as was HLANG. Derived from item # 12d.

G. Number of years of formal education in which English was the language of instruction (YEARS). Derived from item # 22.

H. Year of birth. (BIRTH). Derived from item # 1.

I. Grade in school (GRADE). Derived from item # 21.

J. Highest year of formal education attained by the head of the child's household. (PARENT). Derived from item # 6 of the Household Information Form (see Appendix 16).

Adult MELP Variables. Most of the MELP variables for adults were identical to those used for children as defined above. In particular WHEN, SPEAK, UNDERSTAND, FRIEND, HLANG, YEARS, BIRTH, and GRADE were the same. SIB was dropped for adults because many adults either did not have living siblings or they talk with them only very rarely. PARENT, of course was also dropped.

Three new variables were added: (a) INCOME was taken from the Household Information Form. It asked: "What was the total income of this family during the past year? (This includes wages and salaries, net income from business or form, pension, dividends, interest, rent, social security payments, and any other money income received by members of this family.)" The response alternatives were:

1. $0 - 4,000
2. $5,000 - 9,999
3. $10,000 - 14,999
4. $15,000 - 19,000
5. $20,000 and over
6. Don't know.

(b) NEWS was taken from question #14a of Figure 1. It asked "How often does . . . . read an English newspaper?" The alternatives were "Often", "Occasionally", and "Not at all", and were scored 1, 2, and 3 respectively.

(c) KID was taken from question 12e. It asked for the language normally spoken with children in the household. "English" was scored as 3, any other language as 1, and no response as 2.

The treatment of missing data. In any data collection there will be some protocols which have missing or unusable data for some variables. The reasons for missing data are many. They include refusal or inability of the respondent to answer the question, failure of the interviewer to ask the question or to record the response, and errors in the procedures by which the data are transferred from the questionnaires to computer-readable tapes. For some MELP variables, missing data for an individual respondent caused all of the data from that respondent to be dropped

from the analysis; however, for SPEAK, UNDERSTAND, FRIEND, SIB, and HLANG, a value was substituted (see above) if data were missing. In the case of SPEAK and UNDERSTAND, missing data were coded as "2" ("Just a little") since it was a popular option, and we assumed that missing data on these items were more likely to occur for respondents who were not proficient in English than for those who were more proficient. For FRIEND, SIB, and HLANG, a middle value was used.

Missing data were extremely rare for these variables in any case. About 4% of the responses to SPEAK and UNDERSTAND were either missing or "don't know", as were about 2% of the responses to FRIEND, SIB, and HLANG. These rates were for adults answering about themselves and the Household Respondent answering about a child. Comparable rates for the Household Respondent answering for another adult in the household were slightly higher (see Chapter IX); however these latter, proxy data were not used in the derivation of the scoring keys.

VI. The Criterion Variables

Major objectives of this study were to select a set of MELP questions and to establish concurrent validity for them by comparing responses to them with other measures of Limited English-Speaking Ability. The point has already been made that although no paucity exists of instruments for assessing English proficiency, there is presently no single, widely accepted such measure on which we could rely to obtain the "true" categorization (LESA or non-LESA) of each individual in the field test. Thus, our position was one of having several different measurement approaches to English proficiency -- all admittedly quite fallible -- against which to develop our MELP. Previous chapters have elaborated on the development of three such criterion measures: school list information, a discrete point test, and a direct observation rating procedure (DORP). The discussions in Chapters I and II indicate that these alternatives cannot be ordered among themselves as being "better" or "worse" measures of LESA, they are simply different from each other, with different strengths and weaknesses. The purpose of this chapter is to define each of these measures in detail as used in this study and to present the relationships among them.

1. The Test

Chapter III described in detail the development of two discrete point tests, one for children (younger than 14) and one for adults (14 and older). This section reports the preliminary statistical analyses performed on those tests.

To review briefly, each test was composed of three subtests, one of aural comprehension, one of oral production, and one of oral communication. The children's test was composed of 47 items and 57 possible points. The means and standard deviations of the test scores (total points obtained) for each ethnic-linguistic group of children were as follows:

Group	Sample size	Mean	Standard Dev.
Cubans (Dade Co.)	317	45.0	16.6
Chicanos (El Paso)	364	42.2	16.1
Chinese (S.F.)	146	47.5	12.9
Other Asian (S.F.)	133	54.3	8.6
Navajo (Arizona)	260	52.0	12.3
Overall	1220	47.0	15.1

The comparable information for adults was:

Group	N	Mean	Standard Dev.
Cubans	272	18.8	13.7
Chicanos	202	14.7	12.8
Chinese	111	24.4	17.3
Other Asians	116	39.6	11.2
Navajos	214	39.9	13.7
Overall	915	26.1	17.3

Although the means order themselves similarly across the groups the range of adult means is considerably greater than the range of child means. Generally speaking, the Spanish speakers scored relatively low on the tests while the Other Asians and Navajos scores quite high.* The Chinese showed an intermediate degree of proficiency with the adults having a particularly large amount of within-group variability.

Although these tests were made up of three subtests each, the requirement was for a single, global measure of English proficiency rather than three measures. Two alternatives suggested themselves: the first was to simply use the total number of points scored on the test as an individual's score and assume that the test in fact measured a single dimension interpretable as English proficiency. This was what was done in early analyses of the data, including those described in Chapter V. The second approach was to empirically explore the dimensionality of the test and to construct a unidimensional score for each respondent by weighing the scores of the items in differential ways. This avenue was explored through using principal

components factor analysis. Factor analysis is a general statistical technique which analyzes the co-variation of a number of variables constructed to be mutually uncorrelated with each other. In the present application, if the test actually measured only a single, unidimensional construct (i.e., English proficiency), a single new variable (called a factor) should emerge which was much more prominent than the others, and with which most or all of the original test items would be highly correlated. To the extent that one or more less important and independent factors were found to exist, they would be evidence that the test's total score measures more than simply English proficiency (e.g., IQ, chronological age). A "purified" (i.e., unidimensional) measure of English proficiency could then be constructed by computing a "factor score" for each individual. This factor score is computed by adding the item scores after they have been weighted (multiplied) by coefficients derived by the principal components procedure.

The factor analysis was done separately for children and adults, pooling the 1220 children's test data into a single sample and doing the same for the test data of the 915 adults. All computation was done using the SPSS principal components procedures (Nie, et al, 1975).

Children's Analysis. Each item of the children's test was entered as a variable in the analysis and principal components were taken of the 47 X 47 inter-item product-moment correlation matrix.

Following the usual convention, as described in the SPSS handbook (Nie, et al, 1975, p. 493), only components (factors) with eigenvalues greater than 1.0 were retained. Those eigenvalues are listed below:

factor	eigenvalue	percent of variance
1	16.26	34.6
2	3.36	7.1
3	1.34	2.8
4	1.19	2.5
5	1.03	2.2

* In fact, the children's test was too easy for the Other Asians. It seems likely that there was a definite ceiling effect for some members of that group.

The magnitudes of the eigenvalues corresponding to the various factors indicate their relative importance in terms of variance of the original variables accounted for. Together the five factors accounted for 49.3% of the total variance in the correlation matrix. These five factors were then rotated using a quartimax procedure. The rotated factor matrix is given in Table 1. The entries in this matrix are the correlations of the test items with the various factors and are called "factor loadings."

Table 1: Principal Components analysis of the children's test data; quartimax rotation.

Item	Subtest	F1	F2	F3	F4	F5	h^2
1.	Comprehension	41	-10	42	-14	06	37
2		36	-23	54	05	05	48
3		49	-04	04	-26	22	36
4		32	-15	45	-12	17	38
5		43	-18	37	08	-11	38
6		40	04	-04	-44	06	36
7		-03	-07	09	54	71	80
8		41	05	22	-32	20	37
9		37	-26	42	-18	-11	42
10		49	-12	-02	-01	18	29
11		33	-07	18	-16	-16	20
12	Comprehension	53	-13	08	-03	28	39
13	Production	67	-37	01	13	-03	61
14		70	-13	-19	-15	-01	57
15		71	-15	-19	-08	-04	57
16		69	-42	-07	15	-14	70
17		72	-32	-07	11	-13	65
18		67	-33	-03	12	-15	60
19		73	-24	-07	-05	-04	60
20		60	01	02	-22	04	41
21		65	-07	-07	-14	20	49
22		73	-08	-16	-14	01	59
23		68	-12	-03	-11	12	50
24		70	-02	-14	-04	05	51

Table 1 continued.

Item	Subtest	F1	F2	F3	F4	F5	h^2
25	Production	67	-30	-10	09	-08	57
26	↓	71	-21	-06	-01	09	56
27		66	-21	-06	03	-00	49
28		65	-22	-21	04	06	52
29		66	-15	-05	-13	17	51
30		66	-04	-07	-09	14	47
31		64	-30	-10	13	-02	53
32	↓	64	-33	-06	23	-09	58
33	OCT	48	37	09	-02	-01	37
34	↓	52	26	01	16	-00	37
35		53	29	-04	06	-07	38
36		51	32	02	-01	04	37
37		57	34	08	13	-03	46
38		59	38	02	-05	05	50
39		67	32	-00	16	-07	59
40		58	39	-03	09	02	50
41		68	38	01	15	-14	65
42		64	39	01	05	-01	56
43		48	39	02	-05	04	39
44		63	39	01	08	-04	56
45		58	36	-01	09	-06	48
46		61	46	07	04	-06	59
47	↓	65	42	03	07	-04	61

The last column designated h^2, contains the sums of the squares of the loadings in each row. h^2 can be interpreted as the percentage of each variable's variance participating in the five factors. That these numbers are relatively low indicates either that the items had a high degree of singular variation or were relatively unreliable.

F1 seems to be a general English proficiency factor. It accounts for almost five times the variance of the second factor and all but one item loads on it with a loading greater than 0.3. F1 seems to be anchored most directly by the production items. Thus, it seems clear that it represents the construct that we sought to measure.

F2, accounting for 7.1% of the total variance, is of little substantive interest. The product moment correlation of the loadings in the F2 column with the difficulties of the items is -0.88; thus, this factor should be considered to merely represent item difficulties and be essentially devoid of substantive interest. F3 and F4, representing 2.8 and 2.5 percent of the variance respectively, seem to involve primarily the comprehension subtest. The six highest loadings on F3 are all on items in that test, as are the four highest loadings on F4. A more extensive interpretation of these factors is not obvious. F5 again involves the comprehension items with its primary anchor being item #7 and little else loading on it.

Given the highly dominant first factor in this solution along with the presence of several minor factors which were apparently either unrelated to the content of the test or relatively uninterpretable, the decision was made to use each child's score on the first factor as his test score. Thus, factor scores corresponding to F1 were computed for all children and these were then used in all subsequent analyses as representing the children's performances on the test.

These factor scores will be referred to as FCTR; FCTR is scaled with a mean of zero and a standard deviation of one (over the entire sample of scores).

Adults' Analysis. The 41 items in the adult test were entered as variables into a principal components analysis. As with the child analysis, components with eigenvalues greater than 1.0 were retained and rotated using the quartimax procedure. There were four such components (factors) and their relative importance can be described by the sizes of their respective eigenvalues:

Factor	eigenvalue	Percent of total variance
1	18.7	45.5
2	2.0	5.0
3	1.8	4.3
4	1.2	2.8

Together, the four factors represented 57.6% of the total variance of the 41 items. The rotated factor matrix is given in Table 2, together with the means of the items and the h^2 corresponding to each item.

The factor structure has some similarity to the structure found for the children's test. In both cases the production test appeared to anchor the first factor while the comprehension test showed the weakest properties. In the ACT the average h^2 was lower than in either the APT or the OCT, indicating the likelihood that its items were of lower reliability. This conclusion is reinforced by the pattern of ACT means. All except one fall between .43 and .53. This is particularly significant when one considers that the ACT items were all two-choice, and thus would have expected means of .50 if all responses were randomly made. Therefore respondents did somewhat poorer than chance on the test as a whole. The ACT items appear to load both on F1 and F2; however, they load more highly on F1 (mean loading=.40) than on F2 (mean loading=.28).

F1 is clearly interpretable as a general English proficiency factor. Its variance is over nine times the variance of the next most important factor, and the great majority of the items in the test (37 out of 41) were principally identified with it. Therefore, the factor score corresponding to the first factor was computed for each respondent, and this score was used in all subsequent data analyses as that individual's test score. Conceptually, the factor score (referred to, as in the children's analysis, as FCTR) can be thought of as a purer measure of the central construct under investigation than is the raw total number of points obtained. However, in this particular case, there was little real choice between the two measures, since in the total sample they correlated .986 and in no ethnic group did they correlate less than .973. As in the case of the child factor scores, the adult FCTR scores were standardized over the entire sample with a mean of zero and a standard deviation of one.

Table 2: Principal Components of Adults' Test Data. Quartimax Rotation. (all numbers given to two decimal places, decimal points deleted.)

Item	Subtest	Mean	F1	F2	F3	F4	h^2
1		53	31	41	28	-03	35
2		46	20	35	37	22	35
3		53	46	25	23	18	36
4		48	32	01	-00	54	41
5	ACT	47	49	38	24	-11	45
6		43	53	40	23	-10	51
7		44	46	26	34	-12	41
8		25	00	41	53	02	45
9		47	61	-00	-25	07	44
10		44	57	36	15	-21	52
11		97	76	11	-13	15	63
12		106	80	08	-10	14	67
13		103	78	11	-11	18	67
14		98	78	08	-14	23	69
15		109	81	07	-12	20	72
16		98	78	09	-11	18	66
17	APT	96	76	07	-13	-11	61
18		74	80	12	-23	-01	70
19		64	71	23	-23	-05	61
20		70	77	15	-24	-27	75
21		72	75	15	-22	-29	71
22		63	75	14	-31	-19	72
23		67	73	14	-28	-26	69
24		90	77	09	-16	17	65

Table 2 continued

Item	Subtest	Mean	F1	F2	F3	F4	h^2
25		93	80	06	-13	16	69
26		93	79	07	-15	14	67
27		51	66	-23	18	-05	52
28		45	63	-20	11	09	46
29		47	65	-20	11	03	47
30		45	67	-20	15	-01	51
31		49	67	-25	16	-00	53
32		42	69	-18	10	-14	54
33		57	73	-27	18	07	64
34		48	66	-26	16	03	54
35	OCT	57	77	-29	20	01	71
36		52	74	24	16	-04	64
37		30	59	-09	12	-25	43
38		48	72	-28	12	-10	62
39		49	74	-23	14	-04	62
40		47	75	-21	13	-08	63
41		50	75	-27	17	-10	68

2. The School Lists -

The school list information had two vital strengths relative to its use as a MELP variable.

1. It is very close in definition and purpose to the legislative definition of LESA and can (for some school districts) be directly interpreted as the LEA's way of identifying LESA and non-LESA children.

2. It is inherently categorical rather than continuous in nature and thus provides an excellent guide by which to determine a cut off point on some continuous MELP measure (e.g., a discriminant or regression function).

Unfortunately, however, such school information has one large disadvantage: it is completely locally defined and it is unlikely that any two LEAs will categorize children in just the same way. (This, of course, is a characteristic of the United States' decentralized school system.)

School Lists: Children: The particular school districts from which the present samples were drawn were recommended because they had exemplary screening procedures and/or curricula for children of non-English language backgrounds; but each used its own procedure for determining if a child was to be considered LESA or not. A relatively brief sketch of the procedure used by each school is given below:

A. Dade County Public Schools (Miami): Upon regeristering for the first time in school, each child with a background involving a language other than English (as determined informally by the registration clerk) is usually interviewed by a specialist in the field of English as a Second Language (ESL). As a result of that interview, the child is categorized as non-independent or as independent in English, or, if the results of the interview are not clear cut, he is given in additional assessment in the form of a test -- either the Aural Comprehension Test or the Thumb-nail Test (both locally developed). An intermediate category contains

children who are not clearly in either the independent or the non-independent categories. Children who are "independent" in English are considered to be able to function independently in a monolingual English school setting without supplementary materials or instruction in another language; this is clearly the concept of being non-LESA as defined legislatively. The categories of "non-independent" and "intermediate" are also clearly LESA according to their definitions. Thus, in all analyses reported here involving school lists in Dade County, independent children were categorized as non-LESA and all others as LESA.

B. El Paso: Children were classified as either Spanish Dominant or English Dominant based on their relative performances on parallel forms of a locally-developed grammar test in English and Spanish. Children scoring at the top of both tests or at the bottom of both tests were not on our lists at all. Classification was made on the basis of the difference between the two test scores. A child scoring higher on the Spanish test than on the English test was categorized as "Spanish dominant", while a child scoring higher in English than in Spanish was categorized as "English dominant". In the present analyses, "Spanish dominant" was equated with LESA and "English dominant" was equated with non-LESA. While it would have perhaps been better from the point of view of the project to simply use scores on the English test to define the lists, this was not how El Paso screened its children, and such scores were not available to us in any case.

C. Arizona: Navajo children were taken from two school districts, Window Rock and Ganado. The districts used very different classification procedures: 1. Window Rock: Although Window Rock does not routinely

classify children on English proficiency, they devised lists for us by using scores from comprehension section of the Gates-McGinitie Reading test. All those scoring below their grade level were placed on the "low" list. Thus, those scoring below grade level were interpreted as being LESAs in the present analysis while those scoring at or above grade level were categorized as non-LESA. Although this categorization procedure was initially considered to be marginally relevant to the LESA concept, subsequent examination of the relationship of the Window Rock lists with the other variables in the study led us to discard the information entirely. Details are given in Appendix 5.

2. Ganado: Ganado relied mainly on teacher ratings, but also used the same Thumbnail test (10 completion items) that was used in Miami. Ganado had three categories labeled non-Independent, intermediate, and Independent. Their meanings appeared to be the same as in Miami, and they were interpreted the same as were the Miami lists relative to LESA and non-LESA.

D. San Francisco: San Francisco's classifications were apparently made by the child's teacher after a few weeks of school in the fall. No formal assessment procedure was followed. The classification was dichotomous with categories labeled limited English, and non-limited English. It should be pointed out that these lists were at least 9 months old when our data were gathered. All other sites had updated their classifications of the children within the three months previous to our data collection. It should also be pointed out that all children in the San Francisco sample were selected from the rosters of regular elementary schools and not from the "Newcomers'" or "Education" centers where many new arrivals spend

their first months in the U.S. Thus, it is likely that our San Francisco sample did not include some of the most limited children.* That is, all children in our group knew at least enough English to be judged able to survive in a regular English-language school.

School List Information: Adults. Since primary emphasis for developing a MELP was on children between the ages of 5 and 17 (see Chapter I), field test sites were chosen primarily on the basis of availability of school lists for children and only secondarily on the basis of school list availability for adults. As a consequence, such information was only obtainable for adult samples in Dade County and El Paso and not for adults in Arizona and San Francisco. Therefore, lists could be used as a criterion variable for adults only in Dade County and El Paso. The definitions of the samples are as follows:

School List Information in El Paso - the list information which was available for Chicano adults appeared somewhat suspect for reasons that follow: Upon detailed investigation, CAL discovered that the El Paso lists had not been constructed in any direct way from the results of screening procedures. Rather, they represented current enrollments of individuals in either beginning or advanced ESL classes. Unfortunately, the relation between English proficiency and the level of the class in which the respondent was enrolled appeared to be relatively uncertain. The selection of a particular ESL class by a potential student was always voluntary. Although ESL teachers were available to help people choose the correct class for their ability level, many times the choice was determined by convenience of meeting times, level of the student's aspirations, etc. Given this situation, it would be reasonable to expect that the level of the class in which an individual was enrolled would not be highly related to other indices of the individual's English proficiency -- the MELP questions and FCTR scores in particular.

* We believe that children who are most limited in English proficiency are not difficult to identify with MELP-type questions. It is those children with some English proficiency whose identification is most problematical.

The product-moment correlations between El Paso list placement and the MELP variables are given below and compared with the correlations of the MELP variables and FCTR.

MELP Variable	correlated with LIST	correlated with FCTR
WHEN	.10	.07
SPEAK	.10	.53
UNDERSTAND	.07	.53
KID	.05	.14
FRIEND	.01	.18
HLANG	.04	.08
YEARS	.06	.39
NEWS	-.17	-.32
BIRTH	-.02	.05
GRADE	.03	.19
INCOME	-.08	.15
FCTR	.17	1.00

It can be seen that the correlations of the predictors with FCTR are higher than with list in all but one case. The multiple correlation between all eleven predictors and list was .23 while it was .65 between them and FCTR.

Thus it can be seen that not only was the list information in El Paso not the result of a direct screening procedure for English proficiency, but it also was not related highly to <u>any</u> other measure of proficiency in our study. On this basis list information was discarded for adults in El Paso.

School List Information in Miami - The situation in Miami was quite different. The routine procedure in the Miami adult education program is for each potential ESL student to be interviewed by an ESL specialist when enrolled. A preliminary placement is then made and a follow up interview is conducted three days later to

see if the classification was accurate. Students are encouraged to take tests to help in placing them, but testing is not a required part of the screening procedure.

The following are guidelines for ESL interviewers in making initial placements in Miami:

Beginning Level

1. Understands only limited conversation or none at all
2. Makes errors in using the most frequent grammatical structures
3. Speaks with significant distortions of words
4. Uses very limited vocabulary

Intermediate Level

1. Understands everyday speech when speakers choose words carefully or restate ideas
2. Makes significant grammatical errors of interference
3. Speaks with significant distortion of words
4. Gropes for words and often has to rephrase to be understood

Advanced Level

1. Understands nearly everything a native speaker understands
2. Uses English with few grammatical errors
3. Speaks with minor distortions of pronunciation
4. Uses vocabulary comparable to that of native speakers

As can readily be seen, the description of the Advanced level clearly implies no limitation in English while the other two imply limitations of varying degrees; thus the Beginning and Intermediate levels were designated as LESA and the Advanced list as non-LESA. Statistically, this classification scheme was more closely related to the MELP predictors and FCTR than was the El Paso classification. For Miami, the multiple correlation of the eleven MELP variables with list (dichotomized into LESA - non-LESA) was .48 and the correlation of list with FCTR was .51. Therefore on both definitional and statistical grounds, the decision was made to retain the list information in Dade County as a criterion variable.

3. The Direct Observation Rating Procedure (DORP)

As described in Chapter III, a Direct Observation Rating Procedure (DORP) was developed to serve as a criterion measure of language proficiency against which to derive and validate the MELP (in parallel or combination with the school lists and test). Unfortunately, the development effort was completed too late to use the instrumentation in all sites and to properly train all interviewers in its administration. As a result, relatively complete DORP data were gathered only for the Cuban and Chicano groups, and thus the DORP could not be used as a full-fledged criterion variable in the derivation of the MELP. Nevertheless, the purpose of this section is to report analyses of what DORP data were collected, focusing on its relationship to the other two criterion variables (for the Spanish-speaking groups only, of course). Such analyses provide some additional concurrent validity to both the test and the MELP variables in the sense that the DORP represents a method of assessing English proficiency which is not represented in the test and not directly represented in the lists. (Ratings by teachers or other school personnel, on which some lists were based, could be thought of as being somewhat similar to DORP ratings.) Moreover, the DORP does represent a method for assessing language proficiency which is accepted as face-valid by many specialists.

Table 3 indicates the number of DORP ratings made within each ethnic group.

Table 3

Site	Sample Size	DORP Ratings
Cuban children	317	307
Cuban adults	272	262
Chicano children	364	306
Chicano adults	202	153
Asian children (including Chinese)	279	65

Table 3 continued

Site	Sample Size	DORP Ratings
Asian adults (including Chinese)	227	58
Navajo children	260	61
Navajo adults	214	46

Since the number of DORP ratings made were relatively few in Arizona and San Francisco, they were eliminated entirely from the analyses to be reported below and only those involving Spanish speaking respondents were used.

4. Relationships Among Criterion Measures

Table 4 gives the product-moment correlations among test total score, FCTR, List, and DORP for the Chicano and Cuban children and for test total, FCTR, and List for Chinese, Other Asians, Navajos and all children together. It should be remembered that since list is dichotomous, all correlations with List are point biserial coefficients and can thus be expected to be lower in magnitude than the other coefficients (as indeed they are). Table 4 shows what we might expect with three fallible measures of the same construct: that is, the correlations are substantial, but nowhere near unity. Table 5, which gives the corresponding correlations for Cuban adults, yields very similar results.

An alternate way of looking at the relationship between List and FCTR, our two principal criteria, will be given in the last section of this chapter.

Table 4: Intercorrelations of Criterion Measures for Children

A. Cubans (N=307)

	List	Test	FCTR	DORP
List	---	.43	.37	.46
Test		---	.93	.72
FCTR			---	.66
DORP				---

B. Chicanos (N=306)

	List	Test	FCTR	DORP
List	---	.61	.60	.55
Test		---	.93	.71
FCTR			---	.65
DORP				---

C. Chinese (N=146)

	List	Test	FCTR
List	---	.40	.32
Test		---	.86
FCTR			---

D. Other Asians (N=133)

	List	Test	FCTR
List	---	.31	.26
Test		---	.72
FCTR			---

E. Navajos (Ganado Only)

	List	Test	FCTR
List	---	.31	.30
Test		---	.88
FCTR			---

F. Overall (N=1098)

	List	Test	FCTR
List	---	.45	.43
Test		---	.92
FCTR			---

Table 5: Correlations for Cuban Adults: Criterion Measures

	List	Test	FCTR	DORP
List	-	.52	.49	.48
Test		-	.98	.73
FCTR			-	.72
DORP				-

The question has been raised about the degree to which DORP and FCTR combined might make a criterion variable more valid and reliable than either is alone. To obtain an idea of that, each child's FCTR and DORP scores were simply added together after having been standardized within group. The multiple correlation coefficients of these composite variables with the 10 MELP variables were .73 and .82 for Cubans and Chicanos respectively. The corresponding multiple correlations for FCTR alone are .67 and .73 respectively. Thus, within these groups, the use of FCTR and DORP in combination might have been expected to control about 10% more of the variance of the MELP. Had complete DORP data been obtained for all children such a combination would have been employed, resulting in somewhat better performance figures for the scoring keys derived in Chapter VII. Whether the better performance would have been due simply to greater reliability or also to greater validity of the criterion variable it is impossible to say.

With respect to adults, the situation was slightly different. Again, FCTR and DORP scores were both standardized within group and then added together for each individual. As in the above analysis, this composite variable was then used as the criterion in a multiple regression analysis with the MELP variables as predictors. The multiple correlation coefficients were .70 and .64 for the Cuban and Chicano groups respectively. They compare with .69 and .65 respectively when FCTR is used alone. This indicates that for adults little if any additional performance would be gained by a MELP if it were derived using a combination of FCTR and DORP as a criterion. Certainaly, DORP ratings alone would not seem to be superior to FCTR as a criterion -- except possibly on the basis of face validity alone.

Dichotomizing FCTR. Because the objective of this study was to develop a measure of a dichotomous characteristic, it was necessary to convert FCTR from a continuous variable into a dichotomous one before it could usefully serve as a criterion

measure in the derivation of a scoring key for the MELP. This amounted to defining a cutting point on the FCTR scale such that all children having scores below that value would be considered LESA -- as far as test results were concerned -- and all children scoring at or above that value would be considered non-LESA. But how could that cutting point be determined in a non-arbitrary way? Since norms had not previously been computed for this test, there was no way to interpret what a given score meant relative to any known group distributions. Neither was the test constructed to be criterion-referenced, so inspection of the contents of the items did not help to determine what score ranges might be called LESA and non-LESA respectively. The only link from the test to a dichotomy was the fact that the respondents had taken the test and had been classified LESA or non-LESA by schools. The solution employed, then, was to assume that the schools had given us the correct number of children who were LESA in the sample, even if they had not been correct in their categorization of every individual child. (This is equivalent to assuming that the schools made as many false positive diagnoses of LESA as they did false nagatives.) The cutting point on the test was then determined by placing it such that the same number of children (approximately) were characterized as being LESA by the test as by List. For example, among Cubans, there were 210 children on the LESA school lists out of 317 children. The FCTR cutting point was chosen for Cubans, then, so that the 210 children who scored lowest on FCTR were LESA and the highest 107 were non-LESA. That cutting point was +.45 on the FCTR scale (or approximately 54 in terms of total test points). This procedure was carried out for each group individually and for the entire sample of 1098 as a whole. The FCTR cutting points are given in Tables 6b, 7b, 8b, 9b, 10b, and 11b and ranged from .18 for Chicanos to .63 for Navajos. One way to interpret this range is to ascribe it to differences in the criteria which the schools implicitly or explicitly used in making their

classifications. It may be that a child knowing just enough English to score .30 on FCTR would be assigned to the "English dominant" group in El Paso but to the "limited" group in San Francisco or the "non-independent/intermediate" group in Miami or Ganado. Another interpretation of the differences is that there was a test-culture interaction. Under this interpretation, Chicano children scored systematically lower on the test than did, say, Navajo children even though they had the same English proficiency -- presumably because the test discriminated against Chicanos in non-linguistic ways. Although it is not possible to dismiss the latter possibility, precautions against it were taken by having representatives from all the ethnic groups criticize the test in detail and suggest alternative, more "culture-fair" forms.

It should be noted that there are other possible approaches which could be used in dichotomizing FCTR. One would be to determine a cutting point by examining the contents of the various test items and deciding, in consultation with teachers or other specialists, what mimimum performance would be necessary to consider a person as being LESA. Another would be to choose the cutting point which would minimize the number of individuals for which classification by list and by FCTR disagreed. The former method was not pursued because of the difficulties in arriving rationally at such a cutting point in a non-arbitrary way. The latter method was explored and found to yield results very similar to those of the procedure which was employed.

Adults While the same logic was used in dichotomizing FCTR for adults as was used for children, the procedure was only possible for the Cuban group since that was the only group for which useful list classifications were available. Thus, a cutting point was established only for Cubans and then simply assumed to be valid for the other groups. The cutting point arrived at was +0.1, corresponding to a total test score of approximately 29. When the cutting point of +0.1 was applied to each of

the adult samples, the following numbers and proportions of individuals fell into the LESA and non-LESA categories:

	Overall		Cubans		Chicanos		Chinese		Other Asians		Navajo	
	N	Prop.	N	Prop.	N	Prop.	N	Prop.	N	Prop.	N	Prop.
LESA	444	.49	185	.68	160	.79	56	.50	17	.15	26	.12
non-LESA	471	.51	87	.32	42	.21	55	.50	99	.85	188	.88
Total	915	1.00	272	1.00	202	1.00	111	1.00	116	1.00	214	1.00

Although the overall proportions of LESA and non-LESA individuals are approximately equal within the sample of all adults taken as a whole, the proportions within the ethnic group vary widely -- from 79% LESAs among Chicanos to 12% among Navajos. Therefore, it must be kept in mind that in the present study 78% of all LESAs were Spanish speakers and only 27% of all non-LESAs were Spanish speakers.

5. The Correspondence between List and Dichotomized FCTR

Since both List and FCTR will be used in subsequent chapters as criteria against which to derive scoring keys for the MELP variables, it is important to explore the degree of agreement between these two measures themselves. If they are highly redundant with each other, then it is likely that a given MELP scoring key will yield LESA - non-LESA categorizations which will agree with both criteria or with neither. However, to the extent that the two criteria are themselves not highly correlated, then the possibilities become more complex. The MELP might be more highly in agreement with one criterion than with the other or it might be moderately correlated with both. Given two relatively uncorrelated criteria, a moderate correlation with both would seem preferable since we have already taken the position that the two criteria represent different ways of indexing the LESA - non-LESA

distinction and that there is no consensus that one is "better" than the other. Since the point biserial correlations already reported between List and FCTR were relatively low (.43 for all children pooled and .48 for Cuban adults), we can expect that the correspondence between dichotomized FCTR and List will not be particularly high either.

Table 6 gives the four-fold tables of classification for children in each ethnic-linguistic group. The frequencies in the upper-left and lower-right cells of each table represent individuals for whom list classification and dichotomized FCTR classification agreed, while the frequencies in the lower-left and upper-right cells represent disagreements between the two systems. "% agreement" is the sum of the agreements over the total number of individuals in the Table. An inspection of these numbers immediately confirms our expectations, that the degree of association between these two measures, although substantial, is not as high as would be desired for alternative criteria to be used in the derivation of a single measure. Also, the agreement is substantially higher for the two Spanish speaking populations than for the other groups. These considerations must be kept in mind throughout the presentations in Chapters VII and VIII.

Table 6: Agreement between dichotomized FCTR and School List.

A. Cubans

FCTR (cut pt. =.45)	List: LESA	List: non-LESA	Total
LESA	166	43	209
non-LESA	44	64	108
Total	210	107	317

73% agreement

B. Chicanos

FCTR (cut pt. =.18)	List: LESA	List: non-LESA	Total
LESA	161	29	190
non-LESA	30	144	174
Total	191	173	364

84% agreement

Table 6 continued.

C. Chinese

FCTR (cut pt. =.41)	List: LESA	List: non-LESA	Total
LESA	67	25	92
non-LESA	26	28	54
Total	93	53	146

65% Agreement

D. Other Asian

FCTR (cut pt. =.54)	List: LESA	List: non-LESA	Total
LESA	29	29	58
non-LESA	24	51	75
Total	53	80	133

60% Agreement

E. Navajos (Canado only)

FCTR (cut pt. =.63)	List: LESA	List: non-LESA	Total
LESA	69	25	94
non-LESA	26	18	44
Total	95	43	138

63% Agreement

F. All Children

FCTR (cut pt. =.43)	List: LESA	List: non-LESA	Total
LESA	487	153	640
non-LESA	155	303	458
Total	642	456	1098

72% Agreement

VII: Derivation of LESA Categorization Procedures for Children

In Chapter V, a set of ten MELP variables were defined which were the quantified responses to the MELP questions. However, for any given individual these ten variables were a long way from a single categorization as being either LESA or not. The subject of this chapter is the development of "scoring keys" by which a child can be assigned to the category LESA or not on the basis of his or her values on the MELP variables. Two approaches were taken. The first was to use discriminant analysis. This procedure combines a set of discriminating variables (the MELP variables) in a linear discriminant function such that the resulting composite variable maximally discriminates between the two values of a dichotomous criterion variable (either list or FCTR). The discriminant analysis procedure derives the discriminant function -- which includes a weighting coefficient for each predictor variable -- in such a way that the total number of categorization agreements between the discriminant function and the criterion variable is maximized. Conversely, the total number of "errors" of classification made by the discriminant function relative to the criterion are minimized. The second approach to a scoring key was simply to postulate explicit operational definitions of the LESA and non-LESA categories in terms of the MELP variables and then test the agreement of these definitions against the LESA and non-LESA categories as defined by one or another of the criterion variables. Each of these approaches will be explored in turn.

1. The Evaluation of MELP-Based Definitions of LESA and non-LESA.

Any categorization procedure based on the MELP variables, be it a discriminant function or simply an ad hoc definition, when compared with the categorization of the same respondents by either List or FCTR*, yields a four-fold table which

---

* In this chapter, "FCTR" always means "dichotomized FCTR." See Chapter VI for details.

characterizes the amount of correspondence between the two systems. Such four-fold tables and statistics derived from them will form the basis of our evaluations and comparisons of various possible scoring keys. Consider Table 1 below:

Table 1:

	Criterion Categorization (assumed correct)		
	LESA	non-LESA	Total
MELP-based Categorization	A	B	A + B
	C	D	C + D
	A + C	B + D	A+B+C+D

Such a table compares categorization by discriminant function with categorization by criterion. If A,B,C,D represent the frequencies in the above cells, A and D represent those in the total sample which are categorized the same by both the criterion and the discriminant function. Clearly, the larger A + D, the more effective is the discriminant function in predicting the "correct" categorizations of the individuals in the sample. On the other hand, for the purposes of this study, the crucial objective of a scoring key is to correctly estimate the proportion of LESAs in a population. This is not necessarily the same as minimizing the total number of errors of classification. To achieve the former objective, the frequencies in cells B and C must be roughly equivalent to each other or balanced, while to achieve the latter, B + C is minimized. Thus, it is not necessarily the case that a discriminant function will produce the same marginal frequencies (i.e., A + B and C + D) as the criterion categorizations (A + C and B + D) even for the data set from which it was derived.

In evaluating the performance of any scoring key, two kinds of indices are important: one which measures the accuracy of the scoring key in terms of proportion categorized the same by scoring key and by criterion measure. This will be referred to as "% categorized the same by criterion and MELP." It will equal $(A + C)/(A + B + C + D)$. The other measure is of the agreement between the proportions identified as LESA by the criterion and by the scoring key. It is the difference between the two proportions divided by the latter proportion. In terms of Table 1, it is $(B - C)/(A + C)$ and will be denoted as "% bias". Negative values indicate that the scoring key underestimates the number of LESAs while positive values indicate overestimation.

2. Discriminant Analyses: Child Data

Two discriminant analyses were performed on the data from each ethnic group, one using school list as the criterion and the other using FCTR. Such analyses were done separately for each of the five ethnic groups and also for all groups pooled into a single sample. In all cases the same ten MELP variables were used as discriminators. All analyses were done using the SPSS system.

Table 2 gives the overall accuracy of classification of each discriminant function relative to its particular population and its particular criterion. Accuracy is expressed both as the percent of the group classified in the same category by both the discriminant function (MELP) and the criterion and in terms of the disparity between the proportions classified as LESA by both (%bias).

Tables 3 - 8 give the actual cross-tabulations of classifications by each procedure (criterion vs. MELP) for each group. Percentages in each cell represent percent of the column. For example, in table 3a, 497 children were categorized LESA by both List and MELP, 145 were categorized LESA by List and non-LESA by MELP, etc. Of the 642 categorized LESA by List, 497 of them constitute 77% while 145 make up the remaining 23%. Tables 9 and 10 give the discriminant functions used in the MELP categorizations of Tables 3 - 8. The functions in Table 9 define the MELPs used in Tables 3a, 4a, 5a, 6a, 7a, and 8a while those in Table 10 define the MELPs in Tables 3b, 4b, 5b, 6b, 7b, and 8b.

It is clear from Table 2 that while List and FCTR are different from each other (see Chapter V), the MELP variables predict to each with relatively equal accuracy.

TABLE 2: Performance of discriminant functions derived within each group and across groups when used to classify the same group. Children's data.

	Overall		Cubans		Chicanos		Chinese		Other Asians		Navajos	
	LIST	FCTR	LIST	FCTR	LIST	FCTR	LIST	FCTR	LIST	FCTR	LIST	FCTR
% classified the same by Criterion and MELP	77	78	78	75	87	85	75	73	73	71	70	79
% classified LESA by Criterion	58	58	66	66	52	52	64	64	40	40	69	68
% classified LESA by MELP	55	56	58	58	57	54	55	58	38	44	57	59
% Bias	-5	-4	-12	-13	+10	+ 4	-13	-9	-4	+9	-18	-14

Table 3: Overall Sample: Accuracy of overall discriminant functions.

A. LIST as criterion

	LIST		
	LESA	-LESA	Total
LESA	77%	24%	
	497	111	608
MELP (Discr. funct.)			
-LESA	23%	76%	
	145	345	490
Total	642	456	1098

B. FCTR as criterion

	FCTR (cut=.43)		
	LESA	-LESA	Total
LESA	79%	24%	
	504	108	610
MELP (Discr. funct.)			
-LESA	21%	76%	
	136	350	488
Total	640	458	1098

Table 4: Cubans: Accuracy of Cuban discriminant functions.

A. LIST as criterion

	LIST		
	LESA	-LESA	Total
LESA	77%	21%	
	162	23	185
MELP (Discr. funct.)			
-LESA	23%	79%	
	48	84	132
Total	210	107	317

B. FCTR as criterion

	FCTR (cut=.45)		
	LESA	-LESA	Total
LESA	75%	25%	
	156	27	183
MELP (Discr. funct.)			
-LESA	25%	75%	
	53	81	134
Total	209	108	317

Table 5: Chicanos: Accuracy of Chicano discriminant functions

A. List as Criterion

	LIST		
	LESA	Non-LESA	Total
LESA	87%	18%	
	166	30	196
MELP (Discr. funct.)			
Non-LESA	13%	82%	
	25	143	168
Total	191	173	364

B. FCTR as Criterion (cut = .18)

	FCTR		
	LESA	Non-LESA	Total
LESA	87%	18%	
	166	31	197
MELP (Discr. funct.)			
Non-LESA	13%	82%	
	24	143	167
Total	190	174	364

Table 6: Chinese: Accuracy of Chinese discriminant functions

A. List as criterion

	LIST		
	LESA	Non-LESA	Total
LESA	74%	23%	
	69	12	81
MELP (Discr. funct.)			
Non-LESA	26%	77%	
	24	41	65
Total	93	53	146

B. FCTR as criterion (cut = .41)

	FCTR		
	LESA	Non-LESA	Total
LESA	75%	30%	
	69	16	85
MELP (Discr. funct.)			
Non-LESA	25%	70%	
	23	38	61
Total	92	54	146

Table 7: Other Asians: Accuracy of Other Asians discriminant functions.

A. List as criterion

MELP (d.f.)	LIST: LESA	LIST: -LESA	Total
LESA	64%	21%	
	34	17	51
-LESA	36%	79%	
	19	63	82
Total	53	80	133

B. FCTR as criterion (cut=.54)

MELP (d.f.)	FCTR: LESA	FCTR: -LESA	Total
LESA	68%	28%	
	36	17	53
-LESA	32%	72%	
	22	58	80
Total	58	75	133

Table 8: Navajos: Accuracy of Navajo discriminant functions (Ganado only)

A. List as criterion

MELP (d.f.)	LIST: LESA	LIST: -LESA	Total
LESA	69%	28%	
	66	12	78
-LESA	31%	72%	
	29	31	60
Total	95	43	138

B. FCTR as criterion (cut=.63)

MELP (d.f.)	FCTR: LESA	FCTR: -LESA	Total
LESA	78%	18%	
	73	8	81
-LESA	22%		
	21	36	57
Total	94	44	138

Table 9 : Results of Discriminant Analysis; List as criterion
Showing Standardized (S) and Unstandardized (U) Discriminant Function Coefficients

	CUBAN		CHICANO		CHINESE		OTHER ASIAN		NAVAJO		OVERALL	
SAMPLE SIZE	317		364		146		133		138		1098	
VARIABLES	U	S	U	S	U	S	U	S	U	S	U	S
WHEN	-14	-11	-17	-10	19	17	67	60	-103	-09	-12	-10
SPEAK	-39	-47	-15	-18	50	57	-35	-32	-67	-80	-29	-35
UND	-02	-02	-15	-19	-15	-17	58	51	39	45	-08	-10
SIB	-04	-03	-22	-22	11	11	21	18	10	09	-20	-19
FRIEND	-10	-09	01	01	22	19	27	13	-52	-51	-16	-15
H-LANG	-24	-06	-50	-49	30	23	15	15	-14	-13	-37	-33
YEARS	-31	-54	-01	-01	22	29	10	17	-04	-08	-09	-15
BIRTH	-03	-06	-05	-06	-03	-05	-21	-43	-19	-48	-03	-07
GRADE	-06	-12	-03	-04	-07	-13	-18	-33	-15	-30	-02	-03
PARENT	-04	-05	-02	-03	15	24	11	17	-34	-40	-07	-11
constant	507		650		-163		969		1910		561	
Cutting Point	-19		-04		16		-11		-19		-10	

All numbers given to two decimal places. Decimal points omitted

Table 10: Results of Discriminant Analysis; FCTR as criterion
Showing Standardized (S) and Unstandardized (U) Discriminant Function Coefficient

	CUBAN		CHICANO		CHINESE		OTHER ASIAN		NAVAJO		OVERALL	
SAMPLE SIZE	317		364		146		133		138		1098	
VARIABLES	U	S	U	S	U	S	U	S	U	S	U	S
WHEN	25	20	-11	-07	-09	-07	02	02	294	25	01	01
SPEAK	29	35	-26	-32	39	45	-33	-30	-52	-62	-22	-27
UND	15	17	-04	-06	-60	-66	75	66	30	34	-11	-13
SIB	11	10	-04	-04	-21	-20	35	24	-09	-08	-13	-13
FRIEND	07	07	-17	-17	16	14	-03	-01	-07	-07	-07	-07
H-LANG	-12	-03	-45	-44	-28	-22	12	12	-39	-36	-42	-37
YEARS	22	38	-15	-13	-05	-06	32	53	12	25	-18	-30
BIRTH	-17	-39	-01	-01	36	62	-23	-47	05	12	09	19
GRADE	-05	-09	-02	-04	-04	-07	-15	-28	-27	-56	-01	-03
PARENT	-03	-04	-04	-14	-08	-12	03	06	-18.	-22	-08	-13
constant	903		371		-22 31		11 89		-8 83		-282	
Cutting Point	-17		-03		-13		-10		-21		-10	

All numbers given to two Decimal places. Decimal points omitted

Scoring Key

For each of the two types of discriminant analyses discussed above six separate scoring keys were derived: one for each of the five ethnic groups and a sixth for all the groups combined. Each scoring key was simply a linear equation with the terms being the MELP variables and the coefficients being the unstandardized coefficients given in Tables 9 and 10. Such equations yield a single value for each individual. If that value is above the cutting point (see Tables 9 and 10), the individual is in one category, if it is below the cutting point he is in the other. For example, consider the discriminant function for the Cuban children. It is:

Y=-.14*WHEN-.39*SPEAK-.02*UNDERSTAND-.04*SIB-.10*FRIEND .24*HLANG-.31*YEARS-.03*BIRTH-.06*GRADE-.04*PARENT +5.07

For any Cuban child, if Y is less than -.19, then he or she is categorized as non-LESA. If Y is equal to or greater than -.19, then he or she is LESA.

The five keys for the specific ethnic groups could be used by Census to classify the SIE respondents who are members of these five specific groups as LESA or non-LESA. However, there are many other ethnic groups which were not sampled in this field work. What scoring key should be used to classify these respondents as LESA or non-LESA? One possible scoring key is that derived from the combined data.

To check the accuracy of such a procedure relative to each ethnic group for which data were available, the discriminant functions derived from the combined groups were applied to each respondent's MELP variables to categorize that individual as either LESA or not and these categorizations were compared to the criterion categorizations of both List and FCTR. The results are presented in Tables 11 - 16. Comparing Table 11 with Table 2, it can be seen that between 2% and 4% of accuracy is lost in each group, on the average, when a discriminant function is used which was derived from all 1098 respondents (as opposed to using a

discriminant function derived only on that group). In terms of bias, using the overall discriminant function yields an average absolute percent bias of 18 to 20% while the comparable figure for the locally derived discriminant functions is about 12%. Thus, on an ethnic group by ethnic group basis, using a single discriminant function to categorize all groups resulted in an average decrease of 2% to 4% in the number of respondents categorized the same by MELP and Criterion and an average increase of 6% in the error of prediction of the proportion of LESAs.

Table 11: Performance of overall discriminant function on each ethnic group:

Childrens Data

	Cuban		Chicano		Chinese		Other Asian		Navajo	
	List	FCTR	List	FCTR	List	FCTR	List	FCTR	List	FCTR
% classified the same by Criterion and MELP	75	75	85	82	74	71	68	72	69	67
% classified LESA by Criterion (from table 2)	66	66	52	52	64	64	40	40	69	68
% classified LESA by MELP	72	64	53	60	57	66	20	24	57	43
% Bias	+9	-2	+1	+16	-11	+4	-50	-40	-18	-38

Table 12: Cubans: Accuracy of overall discriminant functions.

MELP	List: LESA	List: -LESA	List: Total
LESA	85%	46%	
	179	49	228
-LESA	15%	54%	
	31	58	89
Total	210	107	317

MELP	FCTR: LESA	FCTR: -LESA	FCTR: Total
LESA	80%	35%	
	167	38	205
-LESA	20%	65%	
	42	70	112
Total	209	108	317

Table 13: Chicanos: Accuracy of overall discriminant functions.

MELP	List: LESA	List: -LESA	List: Total
LESA	86%	16%	
	165	27	192
-LESA	14%	84%	
	26	146	172
Total	191	173	364

MELP	FCTR: LESA	FCTR: -LESA	FCTR: Total
LESA	91%	27%	
	173	47	220
-LESA	9%	73%	
	17	127	144
Total	190	174	364

Table 14: Chinese: Accuracy of overall discriminant functions.

MELP	List: LESA	List: -LESA	List: Total
LESA	74%	26%	
	69	14	83
-LESA	26%	74%	
	24	39	63
Total	93	53	146

MELP	FCTR: LESA	FCTR: -LESA	FCTR: Total
LESA	79%	44%	
	73	24	97
-LESA	21%	56%	
	19	30	49
Total	92	54	146

Table 15: Other Asians: Accuracy of overall discriminant functions.

		List		
		LESA	-LESA	Total
MELP	LESA	36%	10%	
		19	8	27
	-LESA	64%	90%	
		34	72	106
	Total	53	80	133

		FCTR		
		LESA	-LESA	Total
MELP	LESA	45%	10%	
		24	8	32
	-LESA	55%	90%	
		29	72	101
	Total	53	80	133

Table 16: Navajos (Ganado only): Accuracy of overall discriminant functions.

		List		
		LESA	-LESA	Total
MELP	LESA	68%	30%	
		65	13	78
	-LESA	32%	70%	
		30	30	60
	Total	95	43	138

		FCTR		
		LESA	-LESA	Total
MELP	LESA	57%	11%	
		54	5	59
	-LESA	43%	89%	
		40	39	79
	Total	94	44	138

3. Contingency Table Analysis and the Derivation of Explicit Operational Definitions of LESA and non-LESA.

Two sorts of problems attend any attempt to use discriminant analysis to produce a scoring key in the present project. The first is that it was impossible to satisy the statistical assumptions of this sort of analysis. Two such assumptions are that the predictor variables are measured in an error-free way and that they are continuous. The second problem is in the nature of the scoring key produced by such methods. It is a linear equation which adds all predictor variables in a weighted fashion into a single, continuous composite variable with a cut off point to define the categories LESA and non-LESA. Such a scoring key is totally baffling to someone not familiar with multivariate analysis and not readily interpretable even to those who are familiar with it. One of the common questions asked by people attempting to understand how the MELP works is "What patterns of answers to the questions identify a person as a LESA?" That is a fair question, but quite unanswerable within the regression-discriminant analysis context. This section describes the derivation of a scoring key that provides a ready answer to the question. It seeks to enumerate exactly those response patterns (to the MELP questions) defining the LESA category and those defining the non-LESA category. The analysis consisted of two steps: the first involved reducing the number of possible response patterns of the 10 MELP variables to a workable number (from the over 30,000 possible patterns implied by the definitions in Chapter V); and the second was to display the data in appropriately detailed contingency tables so that the effectiveness of various definitions of LESA and non-LESA could be determined.

Reduction of the number of predictor variables

Three strategies were used in reducing the number of possible response alternatives to a manageable size: elimination of relatively redundent predictors,

reduction of the number of possible values of a given MELP variable, and viewing several variables as a composite predictor.

As a first step, consider the inter-correlations of the ten predictors together with List, and FCTR, as computed across all children (Table 17).

Table 17: Product-Moment Correlations: All Sites*

	2	3	4	5	6	7	8	9	10	11	12
1. WHEN	20	19	25	14	36	06	22	-24	-00	22	14
2. SPEAK		83	48	46	45	32	-17	14	30	50	55
3. UNDERSTAND			44	45	42	28	15	13	29	46	53
4. SIB				51	62	16	-00	01	32	45	42
5. FRIEND					37	17	-04	04	29	38	47
6. HLANG						14	-00	-02	32	46	37
7. YEARS							-68	68	03	19	42
8. BIRTH								-82	10	03	-28
9. GRADE									-06	03	26
10. PARENT										27	23
11. LIST*											42
12. FCTR											

* N=1220 for all correlations not involving list. All correlations involving list are based on N=1098, the Window Rock data being excluded. $r > .10$ significant at $p < .01$

Variables with relatively low correlations with the criteria would be early candidates for elimination. Such is the case with WHEN, BIRTH, GRADE, and PARENT. Two highly redundent variables were SPEAK and UNDERSTAND making them clear candidates for combination or for the elimination of one of them. The latter strategy was discarded because the variables were the two most highly related to the criteria. After examining the crosstabulation of SPEAK by UNDERSTAND by each criterion, it was decided to simply add the two variables to form a single variable with a range of from 2 to 10 which was called SPUND. This reduced 25 SPEAK X UNDERSTAND response patterns to nine.

A second composite variable was formed by combining the three variables based on domains of language use -- HLANG, SIB, and FRIEND. The crosstabulation of the three variables (reproduced below) indicates that they form a three-item Guttman scale (Guttman, 1944).

HLANG=English

FRIEND \ SIB	English	not English
English	333	33
not English	20	10

HLANG=not English

FRIEND \ SIB	English	not English
English	177	26
not English	257	375

The perfect scale types are:

Type 0	Type 1	Type 2	Type 3
HLANG=not English	HLANG=not English	HLANG=not English	HLANG=English
SIB=not English	SIB=not English	SIB=English	SIB=English
FRIEND=not English	FRIEND=English	FRIEND=English	FRIEND=English

94% of all responses were one of these perfect scale types. On the basis of this analysis, the four-position scale USE was defined as the number of responses of "English" given by a respondent to HLANG, SIB, and FRIEND. This reduced 27 possible response patterns to four with very little loss of information.

Finally, WHEN, BIRTH, GRADE, and PARENT were eliminated from the battery of predictors on the basis of relatively low correlations with the criteria and low beta-weights in the multiple regression analysis. (see Appendix 6) This, then, left three predictor variables: SPUND, USE and YEARS with a total of 9 X 4 X 9 or 324 possible response patterns. To further reduce this number, YEARS was treated as having 5 alternatives: 0 or 1, 2, 3, 4, and 5 or more. This resulted in 180

possible response patterns. The product-moment correlations among these variables and with List and FCTR are given in Table 18. Table 19 presents the number of respondents with each possible combination of SPUND, USE, and YEARS values and the percent of them who were categorized as LESA by List. (For example, in Table 20, there were children who had SPUND values in a 2 to 7 range and had a USE value of zero and a YEARS value of zero or one, and 93% of them were LESA as determined by List.)

Table 18: Product-Moment Correlations: All Sites*

	2	3	4	5
1. SPUND	58	31	50	57
2. USE		20	53	51
3. YEARS			19	42
4. LIST*				42
5. FCTR				

---

* See foot note for Table 17.

Cochran and Hopkins (1961) give an algorithm for labeling each cell of such a matrix as being either a LESA cell or a non-LESA cell so as to maximize the total number of correct categorizations. Let p equal the proportion of LESA individuals in the entire population of respondents -- in this case $p = \frac{642}{1098}$ or .58. Then, if the proportion of LESAs in any given cell equals or exceeds that number, the cell is labeled as a LESA response pattern.

Table 19: Percent LESA children by List for each combination of SPUND, USE and YEARS. ( ) denotes n in cell.

YEARS = 0 or 1

SPUND	USE 0	1	2	3
2 - 7	.93 (155)	.92 (52)	.76 (17)	.50 (4)
8	.73 (26)	.81 (42)	.54 (24)	.16 (19)
9	.60 (5)	1.00 (7)	.42 (12)	.0 (14)
10	0 (1)	.25 (4)	.22 (9)	.06 (47)

YEARS = 2

SPUND	USE 0	1	2	3
2 - 7	.98 (63)	.72 (25)	.63 (8)	.50 (2)
8	.78 (18)	.79 (29)	.38 (16)	.16 (19)
9	0 (2)	.50 (6)	0 (1)	.33 (9)
10	1.00 (1)	.50 (8)	.40 (10)	.02 (41)

YEARS = 3

SPUND	USE 0	1	2	3
2 - 7	1.00 (19)	.68 (19)	.67 (15)	1.00 (1)
8	.56 (16)	.60 (16)	.70 (20)	.25 (8)
9	.50 (2)	.67 (6)	1.00 (1)	0 (4)
10	.50 (2)	.33 (3)	.38 (8)	.28 (18)

Table 19 continued.

YEARS = 4

USE

SPUND	0	1	2	3
2 - 7	.88 (8)	1.00 (4)	1.00 (2)	ND
8	.76 (17)	.38 (13)	.46 (13)	.29 (7)
9	ND	0 (2)	.25 (8)	.17 (6)
10	.50 (6)	.60 (10)	.22 (9)	.13 (15)

YEARS > 4

USE

SPUND	0	1	2	3
2 - 7	.83 (12)	.75 (4)	0 (1)	ND
8	.60 (10)	.20 (15)	.09 (11)	.50 (8)
9	1.00 (2)	.50 (6)	0 (3)	.33 (3)
10	.40 (5)	.33 (15)	.33 (12)	.10 (21)

ND= No data in cell

All cells having a proportion of LESAs less than .58 are labeled "non-LESA". In Table 19 all non-LESA cells are underlined. Taking the resulting sets of LESA and non-LESA cells and using them as a scoring key, the following agreement with list categorization is obtained.

		List		
		LESA	-LESA	Total
MELP (p=.58)	LESA	81%	24%	
		523	109	632
	-LESA	19%	76%	
		119	347	466
	Total	642	456	1098

% categorized the same by List and MELP = 79%
% categorized LESA by List = 58.5%
% categorized LESA by MELP = 57.6%
% Bias = -2%

While this compares very well with the performance of the scoring keys derived by discriminant analysis, it is not a face valid definition of LESA and non-LESA. The pattern of non-LESA cells in Table 19 is somewhat irregular, with, for example, several cells having USE=zero being labeled non-LESA while similar cells with higher values of USE are labeled LESA. Such irregularities are probably due to the small number of respondents in many of the cells.

In order to make the definitions of LESA and non-LESA more face-valid, we looked for relatively simple combinations of response patterns that would correspond closely to the cell assignments produced by the above algorithm. For example, consider Definition 1.

Definition 1:

A non-LESA child is one with: USE score of 3 or a $PUND score of 9 or 10 (or both)
A LESA child is one with any other response pattern.

The correspondence of Definition 1 with list is:

	List: LESA	List: -LESA	Total
LESA	83%	33%	
	531	152	683
-LESA	17%	67%	
	111	304	415
Total	642	456	1098

76% Classified the same by List and Def. 1
58% Classified LESA by List
62% Classified LESA by Def. 1
% Bias = +6

The 76% accuracy of this simple rule compares reasonably well with both the 79% maximum accuracy attainable using SPUND, USE, and YEARS, and the 77% and 78% accuracies detained by the discriminant functions (Table 2). The definition overestimates the number of LESAs to a modest extent.

Now consider a slightly more complex definition:

Definition 2:

A non-LESA child is one with at least one of the following patterns:

1. A USE score of 3
2. A SPUND score of 10
3. A SPUND score of 8 or 9 and a USE score of 1 or 2 and a YEARS score greater than 3.

A LESA child is one with any other response pattern.

The correspondence of Definition 2 with List is:

	List		
	LESA	-LESA	Total
LESA	85%	29%	
	543	133	676
-LESA	15%	71%	
	99	323	422
Total	642	456	1098

79% Classified the same by List and Def. 2
58% Classified LESA by List
62% Classified LESA by Def. 2
% Bias = +5

Definition 2 performs slightly better than Definition 1 both in terms of providing a slightly smaller overestimation of LESAs and in terms of classifying more people the same as did List. Since Definition 2 is preferred, its performance by group both using List and FCTR as criterion is given in Tables 20-26. Comparing the performance figures for Definition 2 (Table 20) with those of the overall discriminant function (Table 11), we see overall performance being highly similar with the discriminant functions slightly underestimating the number of LESAs and Definition 2 slightly overestimating them. Performance within group was considerably more variable; however, the same patterns generally emerged. The MELP, regardless of form tends to slightly overestimate the number of LESAs or be quite accurate in the Spanish and Chinese groups, while it rather severly underestimates the number of LESAs in the Other Asian and Navajo groups. The reason for this is not entirely clear. One possible factor is that both dichotomous criteria were geared to the local schools' definitions of LESA and non-LESA. (In all analyses reported above, FCTR was cut at a different place in each group in order to dichotomize it.)

Table 20: Performance of Definition 2 relative to List and FCTR; by group; Children's Data.

	Overall		Cubans		Chicanos		Chinese		Other Asians		Navajos	
	List	FCTR	List	FCTR	List	FCTR	List	FCTR	List	FCTR	List	FCTR
% Classified the same by criterion and Def. 2	79	77	80	74	85	82	75	69	71	71	70	72
% Classified LESA by criterion	58	58	66	66	52	52	63	63	40	40	68	68
% Classified LESA by Def. 2	62	62	72	72	62	62	72	72	29	29	59	59
% Bias	+5	+6	+9	+9	+18	+18	+13	+13	-28	-28	-14	-14

Table 21: Overall Sample: Accuracy of Definition 2.

A. List as criterion

		List		
		LESA	-LESA	Total
Def. 2	LESA	85%	29%	
		543	133	676
	-LESA	15%	71%	
		99	323	422
	Total	642	456	1098

B. FCTR as criterion

		List		
		LESA	-LESA	Total
Def. 2	LESA	83%	31%	
		534	142	676
	-LESA	17%	69%	
		106	316	422
	Total	640	458	1098

Table 22: Cubans: Accuracy of Definition 2.

A. List as criterion

		List		
		LESA	-LESA	Total
Def. 2	LESA	90%	37%	
		188	40	228
	-LESA	10%	63%	
		22	67	89
	Total	210	107	317

B. FCTR as criterion

		FCTR		
		LESA	-LESA	Total
Def. 2	LESA	85%	46%	
		178	50	228
	-LESA	15%	54%	
		31	58	89
	Total	209	108	317

Table 23: Chicanos: Accuracy of Definition 2.

A. List as criterion

		List		
		LESA	-LESA	Total
Def. 2	LESA	95%	25%	
		181	43	224
	-LESA	5%	75%	
		10	130	140
	Total	191	173	364

B. FCTR as criterion

		FCTR		
		LESA	-LESA	Total
Def. 2	LESA	92%	28%	
		175	49	224
	-LESA	8%	72%	
		15	125	140
	Total	190	174	364

Table 24: Chinese: Accuracy of Definition 2.

A. List as criterion

		List		
		LESA	-LESA	Total
Def. 2	LESA	87%	45%	
		81	24	105
	-LESA	13%	55%	
		12	29	41
	Total	93	53	146

B. FCTR as criterion

		FCTR		
		LESA	-LESA	Total
Def. 2	LESA	83%	54%	
		76	29	105
	-LESA	17%	46%	
		16	25	41
	Total	92	54	146

Table 25: Other Asians: Accuracy of Definition 2.

A. List as criterion

		List	
	LESA	-LESA	Total
Def. 2. LESA	49%	15%	
	26	12	38
-LESA	51%	85%	
	27	68	95
Total	53	80	133

B. FCTR as criterion

		FCTR	
	LESA	-LESA	Total
Def. 2 LESA	49%	15%	
	26	12	38
-LESA	51%	85%	
	27	68	95
Total	53	80	133

Table 26: Navajos: (Ganado) Accuracy of Definition 2.

A. List as criterion

		List	
	LESA	-LESA	Total
	71%	33%	
Def. 2 LESA	67	14	81
-LESA	29%	67%	
	28	29	57
Total	95	43	138

B. FCTR as criterion

		FCTR	
	LESA	-LESA	Total
LESA	72%	30%	
Def. 2	68	13	81
-LESA	28%	70%	
	26	31	57
Total	94	44	138

However, if San Francisco and Arizona schools use higher criteria of English ability in order to place a child on the non-LESA list, then the MELP should underestimate the number of LESAs in both the Chinese and Other Asian groups, since both attended San Francisco schools. This is not the case; it overestimates the Chinese and underestimates the Other Asians. Of course, it is possible that the schools systematically demand more English from one group than the other, but that seems unlikely.

Another hypothesis might be that since the average level of English proficiency among other Asians and Navajos (as measured by the Test) was quite high, parents might use different standards of comparison in those groups and systematically underrate their children on the important variables SPEAK and UNDERSTAND relative to parents in the other groups where the general level of English proficiency and use is less. Unfortunately, however, such a tendency would lead to an opposite effect to the one observed -- an overestimation of LESAs in the more proficient group.

A third, less interesting explanation may stem from the different distributions of English proficiency in LESA and non-LESA categories within the various groups. The observed underestimation effect could obtain if most LESA children in the Navajo and Other Asian groups were just below the cut-off point between the two categories (on the test) while most of the non-LESAs were considerably above it in each of those groups. One needs only to assume that misclassification by the MELP is simply a direct function of the distance of the individual's test score from the cut-off point. Similarly, an overestimation of LESAs could occur if most non-LESAs were just above the cut-off score while most LESAs were considerably below it. The within-group test/distributions are generally consistent with this explanation.

4. Scoring Keys to be Recommended for Use with SIE Data

On the basis of the analyses detailed above, three different scoring keys can be recommended as having done the test overall job of replicating the LESA and non-LESA categorizations of the children in the field test -- as categorized by school list and dichotomized FCTR. Two of these scoring keys take the form of linear equations employing the ten MELP variables as terms and multiplying each by a coefficient. One was derived with FCTR as the criterion and the other with list as the criterion. The equations are given below:

Y_{FCTR}=-2.82+.01*WHEN-.22*SPEAK-.11*UNDERSTAND-.13*SIB-.07*FRIEND-.42*HLANG-.18*YEARS +.09*BIRTH-.01*GRADE-.08*PARENT.

Y_{LIST}=5.61-.12*WHEN-.29*SPEAK-.08*UNDERSTAND-.20*SIB-.16*FRIEND-.37*HLANG-.09*YEARS -.30*BIRTH-.02*GRADE-.07*PARENT.

If, for any individual child, the obtained value of Y_{FCTR} is greater than or equal to -.10, then the child is to be categorized as LESA. If the value obtained is less than -.10, the child is non-LESA. Exactly the same rule applies to Y_{LIST}, with -.10 also being the cutting point for that equation.

The third scoring key is Definition 2 in Section 3:

A child is to be considered non-LESA if his response pattern meets at least one of the following conditions:

1. SPUND=10
2. USE=3
3. SPUND= 8 or 9 and USE= 1 or 2 and YEARS greater than 3.

All other children are to be considered LESA.

It is important to stress that these scoring keys have been derived and calibrated for optimal performance on the field test data only. Chapter IX will take up the problems in applying these scoring keys to the SIE data in order to derive

national estimates of LESA individuals. At this point, a simple warning is in order: It is likely that some recalibration will be necessary before these scoring keys can be used to estimate percentages of LESAs from SIE data.

VIII. Derivation of Scoring Keys - Adults

The field test design was considerable different for adults than it was for children. In particular, while all children were sampled from lists provided by local school districts, lists of adults could be obtained from schools in only two locations -- Dade County and El Paso. Thus, in those sites, adult samples were chosen entirely from lists of individuals who were currently enrolled in the local adult education program or who had recently been so enrolled. In the other locations, adults were selected from the households of the children's sample. This difference in sampling strategy probably resulted in more heterogeneity of adults between sites than would have been the case if the same sampling plan had been used in all locations. It is important, then, to describe the samples of adults in somewhat more detail than was necessary for children.

1. Description of Adult Samples

In the Cuban (Dade County) and Chicano (El Paso) samples, respondents were essentially self-selected in the sense that they had enrolled themselves in adult education programs. On the other hand, the adults in the Navajo (Arizona) and Asian (San Francisco) groups were selected on the basis of the elementary school-aged children in their households having been screened for English proficiency and thus placed on a child list. The adult groups differed on many characteristics; but two variables, age and highest educational level attained, are displayed in Tables 1 and 2 as general indices of the differences among the groups. It should be noted that over a third of the Cubans were over 60 years old while no other group had more than 5% over that age. Also, teenagers were relatively numerous only in the Other Asian and Navajo groups. With respect to education, Cubans, Chinese, and Other Asians were much more highly educated than Chicanos and Navajos. Between one-third and one-half of the former groups reported having had at least

some post-secondary education while only 5% and 15% of the latter two groups respectively reported post-secondary work. These differences in demographic characteristics across the ethnic groups must be kept in mind when interpreting the results of analyses.

Table 1: Adults: Per cent of each group in each age category (numbers in parentheses indicate cumulative percentages)

Age	Cubans	Chicanos	Chinese	Other Asian	Navajo
14 - 18	1 (1)	4 (4)	6 (6)	21 (21)	26 (26)
19 - 30	5 (6)	26 (30)	14 (20)	14 (35)	21 (47)
31 - 40	17 (23)	32 (62)	32 (52)	31 (66)	32 (79)
41 - 60	42 (65)	33 (95)	45 (97)	31 (97)	18 (97)
61 and over	35 (100)	5 (100)	3 (100)	3 (100)	3 (100)
Total N	272	202	111	116	214

Table 2: Adults: Highest Grade Reached (%)
(numbers in parentheses indicate cumulative percentages)

Highest Grade	Cubans	Chicanos	Chinese	Other Asian	Navajo
none - 6th grade	21 (21)	65 (65)	19 (19)	8 (8)	19 (19)
7 - 9th grade	19 (40)	19 (84)	18 (37)	10 (18)	29 (48)
10 - 12 grade	24 (64)	11 (95)	31 (68)	32 (50)	37 (85)
College	19 (83)	2 (97)	23 (91)	40 (90)	9 (94)
Graduate Work	16 (99)	3 (100)	9 (100)	10 (100)	6 (100)
Total N	272	202	111	116	214

2. The Analysis Plan for the Adult Data

In general, the analyses for adults were designed to be analogous to those for children. An important difference, however, was that list information was not available for many adults, and so the test became the primary criterion measure of English proficiency. The analyses can be very briefly summarized as follows:

1. A dichotomous criterion variable, interpretable as a categorization of LESA and non-LESA, was constructed as described in Chapter VI.

2. Using this dichotomous criterion variable, discriminant analyses were run. Eleven MELP variables served as discriminators, and separate analyses were run both within and across groups.

3. Contingency table analysis (à la Cochran and Hopkins) was performed using five of the eleven predictors. This led to the construction of an explicit operational definition of LESA-non-LESA which could be used as an alternative to the discriminant function.

3. Discriminant Analysis: Adult Data

The procedure for doing discriminant analysis was generally the same as that with the child data. The SPSS statistical routines were used, and all analyses used the eleven MELP predictor variables defined in Chapter V. The dichotomized FCTR score was used as the criterion variable, and separate analyses were done for each of the five ethnic groups as well as over all groups. For the Cuban group, a separate analysis was done using the list information as the criterion.

The results are presented in Tables 3-5. Table 3 presents the four-fold tables characterizing the degree of success with which the MELP variables could predict LESA and non-LESA categorizations as defined by FCTR. The total percent of correct classifications and the bias are given in Table 4. It should be noted here

that the discriminant functions involved in these analyses were different for each group. That is, for Cubans, the predictions of LESA were made strictly on the basis of the discriminant functions derived from the Cuban data only; for Chinese, the predictions are based on a strictly Chinese discriminant function, etc.

It is clear that, across groups, the percent of individuals classified the same by MELP and FCTR is relatively stable -- between 76% and 84%. However, the amount of bias in predicting the proportion of LESAs in a group varies considerably. In terms of the difference between the percent of LESAs as determined by FCTR and the percent determined by MELP, the range is from predicting 7% too few LESAs among Cubans and Chicanos to predicting 13% too many LESAs in the Other Asian group. But in terms of percent bias (the difference between the two percents divided by the percent LESA as determined by FCTR), the figures range from predicting 11% too few LESAs among Cubans to predicting 88% too many among Other Asians and Navajos.

Table 3: Results of discriminant analysis: accuracy of prediction of eleven MELP variables, using FCTR as the criterion.

A. Cubans

Predicted	FCTR LESA	FCTR -LESA	Total
LESA	77%	26%	
	142	23	165
-LESA	23%	74%	
	43	64	107
Total	185	87	272

B. Chicanos

Predicted	FCTR LESA	FCTR -LESA	Total
LESA	84%	24%	
	135	10	145
-LESA	16%	76%	
	25	32	57
Total	160	42	202

Table 3 continued.

C. Chinese

Predicted	FCTR LESA	FCTR -LESA	Total
LESA	88%	29%	
	49	16	65
-LESA	13%	71%	
	7	39	46
	56	55	111

D. Other Asian

Predicted	FCTR LESA	FCTR -LESA	Total
LESA	71%	20%	
	12	20	32
-LESA	29%	80%	
	5	79	84
	17	99	116

E. Navajo

Predicted	FCTR LESA	FCTR non-LESA	Total
LESA	77%	15%	
	20	29	49
-LESA	23%	85%	
	6	159	165
Total	26	188	214

F. Overall

Predicted	FCTR LESA	FCTR non-LESA	Total
LESA	88%	22%	
	389	102	491
-LESA	12%	78%	
	55	369	424
Total	444	471	915

Table 4: Accuracy of the Within-group-derived discriminant functions, predicting dichotomized FCTR.

	Overall	Cubans	Chicanos	Chinese	Other Asians	Navajo
% Respondents categorized the same by FCTR and MELP	83	76	83	79	78	84
% LESA by FCTR	49	68	79	50	15	12
% LESA by MELP	54	61	72	59	28	23
% Bias	+11	-11	- 9	+16	+88	+88

Table 5:

Results of Discriminant Analysis FCTR as criterion: Showing Standardized (S) and Unstandardized (U) Coefficients
Adult Data:

	Discriminant Function Coefficients											
	Cuban		Chicano		Other Asian		Navajo		Asian		Overall	
Sample Size	272		202		116		214		111		915	
Variables	U	S	U	S	U	S	U	S	U	S	U	S
WHEN	-.12	-.06	.57	.37	-.08	-.05	.53	.07	-.30	-.19	-.08	-.06
SPEAK	-.41	-.42	-.31	-.31	-.62	-.58	-.11	-.12	-.46	-.60	-.25	-.34
UNDERSTAND	-.22	-.25	-.35	-.38	.17	.15	-.22	-.22	.04	.05	-.13	-.18
KID	-.13	-.08	.00	.00	.00	.00	-.21	-.20	.09	.08	-.05	-.05
FRIEND	-.10	-.05	-.15	-.08	.03	.03	-.06	-.05	-.01	-.01	-.05	-.04
HLANG	.79	.20	.38	.24	.07	.07	-.07	-.07	.36	.25	.12	-.10
YEARS	-.04	-.06	-.15	-.33	-.00	-.03	-.04	-.16	-.14	-.69	-.05	-.28
NEWS	.27	.21	.42	.31	.77	.42	.42	.28	-.25	-.19	.19	+.15
BIRTH	-.04	-.06	-.08	-.09	-.39	.51	-.12	-.15	-.04	-.04	-.10	-.15
GRADE	-.07	-.32	-.04	-.15	-.09	-.36	-.07	-.28	-.01	-.03	-.03	-.16
INCOME	-.05	-.05	-.20	-.17	-.01	-.01	.07	-.09	-.13	-.12	-.06	-.06
CONSTANT	1.80		.26		3.49		1.26		2.81		2.06	

Table 5 gives the unstandardized and standardized discriminant coefficients on which the analyses discussed above were based.

Since list information was available for Cubans, it was possible to do a discriminant analysis within that group only using list as the criterion variable. The results of this analysis are presented in Tables 7 and 8.

Table 6: Results of discriminant analysis for Cubans using dichotomized School List as the criterion variable.

	List		
	LESA	non-LESA	Total
LESA	73%	25%	
	135	22	157
Predicted			
-LESA	27%	75%	
	49	66	115
Total	184	88	272

74% classified the same
LESA by list= 68%
LESA by MELP= 58%
Bias=-15%

Table 7: Results of discriminant analysis for Cuban adults showing standardized (S) and unstandardized (U) discriminant function coefficients. School list information as the criterion variable. (All numbers given are to two decimal places. Decimal points omitted.)

Sample Size		272
Variables	U	S
WHEN	03	02
SPEAK	-62	-64
UNDERSTAND	15	17
KID	-01	00
FRIEND	-02	00
HLANG	-69	-17
YEARS	-08	-12
NEWS	28	22
BIRTH	-19	-27
GRADE	-10	-47
INCOME	-01	-01
CONSTANT	300	

Within the Cuban adult sample, the MELP variables do not relate to the lists quite as well as they do to FCTR. They classify slightly fewer individuals the same when list is the criterion than when FCTR is, and they do so with more bias relative to list than relative to FCTR.

Performance of the overall discriminant function by group. Since it will not be possible for NCES to derive a separate discriminant function for each ethnic group surveyed by the SIE, the discriminant function derived from the entire pool of adult field test data must be evaluated as to how well it performs within each

ethnic group involved in the field test. In order to do this, the aggregate analysis reported in Table 3F was broken out by ethnic group. The results are given in Tables 8 and 9. They indicate that the overall discriminant function does reasonably well within each group. In terms of percent respondents categorized the same, the overall function does better in the Chicano, Other Asian, and Navajo groups than do the locally derived functions and it does slightly worse in the Cuban and Chinese groups than do the local functions. In terms of bias, the difference between the percent LESA by FCTR and the percent LESA by MELP ranged from 3% for Other Asians to 14% for Chinese. Expressed as percent, the bias ranges from an underestimate of 18% for Other Asians to an overestimate of 27% and 31% for Chinese and Navajos respectively. These bias figures compare favorably with those deriving from the local discriminant functions given in Table 4. This analysis unequivocally supports the conclusion that, for the ethnic groups represented in this field test, little if anything would be gained by using locally derived discriminant functions instead of using the discriminant function derived from all groups pooled.

Table 8: Accuracy by group of discriminant function derived from entire sample.

A. Cubans

		FCTR	
	LESA	-LESA	Total
LESA	90%	61%	
	167	53	220
Predicted			
-LESA	10%	39%	
	18	34	52
Total	185	87	272

B. Chicanos

		FCTR	
	LESA	-LESA	Total
LESA	97%	60%	
	155	25	180
Predicted			
-LESA	3%	40%	
	5	17	22
Total	160	42	202

C. Chinese

		FCTR	
	LESA	-LESA	Total
LESA	91%	36%	
	51	20	71
Predicted			
-LESA	9%	64%	
	5	25	40
Total	56	55	111

D. Other Asian

		FCTR	
	LESA	-LESA	Total
LESA	53%	5%	
	9	5	14
Predicted			
-LESA	47%	95%	
	8	94	102
Total	17	99	116

E. Navajo

		FCTR	
	LESA	-LESA	Total
LESA	73%	8%	
	19	15	34
Predicted			
-LESA	27%	92%	
	7	173	180
Total	26	188	214

Table 9: Accuracy within each group of the discriminant function derived from entire sample.

	Overall (from Table 4)	Cubans	Chicanos	Chinese	Other Asians	Navajos
% Respondents categorized the same by FCTR and MELP	83	74	85	77	89	90
% LESA by FCTR (from Table 4)	49	68	79	50	15	12
% LESA by MELP	54	81	89	64	12	16
% Bias	+11	+19	+13	+27	-18	+31

4. Derivation of a Scoring Key Through Contingency Table Analysis

An alternative to the discriminant function approach is the direct analysis of a multiway contingency table according to the procedure reported by Cochran and Hopkins (1961). This approach was employed for the child data (see Chapter VII). Instead of deriving a linear equation to categorize individuals, this method simply enumerates all possible patterns of responses to the MELP variables and assigns each to either LESA or non-LESA according to the relative numbers of LESA and non-LESA respondents (as determined by the criterion measure) displaying that particular response pattern. An advantage of the method is that it makes no assumptions about the distributions of the predictors (except for assuming that they are discrete), but a disadvantage is that it becomes unwieldy with a large number of possible response patterns. In order to apply it to the child data, the number of MELP predictors was reduced by elimination and consolidation from ten to three. A similar reduction was also needed in order to apply it to the adult data. The first part of this section, then, will describe the process of reducing the number of predictors to a manageable number and the second will report the analysis proper.

Reducing the number of MELP variables

In order to make the data restricted enough for the Cochran and Hopkins analysis, the number of possible response patterns were reduced. There were three possible ways of doing that:

1. Elimination of variables
2. Reducing the number of response alternatives within a variable
3. Constructing a single composite variable from several variables

In the child analysis, all three strategies were used. That is, WHEN, PARENT, BIRTH, and GRADE were eliminated. The second strategy was employed with YEARS, and

the third was employed in combining SPEAK and UNDERSTAND into SPUND (thus reducing 25 possible response patterns to 9) and in combining HLANG, SIB, and FRIEND into USE (reducing 27 possible response patterns to 4). All three strategies were also used in the adult data.

SPUND - As with the child data, SPUND was defined simply as the sum of the numerical values of SPEAK and UNDERSTAND. The justification for this was as follows: first, the two variables were highly correlated in all ethnic groups (approximately .80 in all groups and overall). Second, both were approximately equally correlated with FCTR and inspection of the three way contingency table of SPEAK by UNDERSTAND by FCTR did not show any distinctive relationships between any two of the three. Thus, a more intricate combining of the two variables did not seem called for. Third, the possibility of eliminating one or the other on grounds of parsimony was not pursued because the two variables were the most closely related to FCTR of any of the predictors, and the inclusion of both was thought to aid the reliability of the MELP.

The USE variables - There were three language use variables among the eleven predictors: HLANG, KID, and FRIEND. These were tested to see if they formed a Guttman scale in the same way that HLANG, SIB, and FRIEND did for children.

The three way crosstabulation of the items is given below:

HLANG= not English

FRIEND \ KID	not English	English
not English	550	48
English	57	42

HLANG= English

FRIEND \ KID	not English	English
not English	29	53
English	9	127

In order for there to be a meaningful Guttman scale, four of the cells must be almost empty and four must be relatively large. Clearly, that is not the case in the above table, and so the idea of compositing these variables was dropped.

In order to decide which variables to eliminate from the set of predictors, two sorts of evidence were inspected. First, we inspected the correlations of each of the eleven predictors with FCTR (not dichotomized) within each group and overall. Those correlations are reproduced below:

	Ethnic Group					
MELP Variable	Cuban	Chicano	Chinese	Other Asian	Navajo	Overall
WHEN	14	07	45	33	-06	41
SPEAK	58	53	74	47	50	73
UNDERSTAND	58	53	69	44	45	71
KID	04	14	53	08	42	41
FRIEND	18	18	56	37	37	49
HLANG	-06	08	54	20	40	45
YEARS	30	39	73	39	59	69
NEWS	-41	-32	-57	-38	-55	-55
BIRTH	16	05	38	32	20	34
GRADE	40	29	44	44	56	43
INCOME	18	15	30	37	15	31

A rank ordering of these correlations indicates that the most important predictors besides SPEAK and UNDERSTAND are YEARS, NEWS, FRIEND, HLANG, and GRADE. These variables were placed in a stepwise discriminant analysis within each group, and the order in which the variables were entered into the analysis was observed. The results indicated that the four most important variables (in addition to SPUND) for predicting dichotomized FCTR were YEARS, NEWS, HLANG, and GRADE. These variables plus SPUND were therefore retained for use in the contingency table analysis.

Even within this reduced set of variables, however, it was important to further reduce the number of possible response patterns. Thus, YEARS, NEWS, HLANG, and GRADE were dichotomized. This was done by going back to the crosstabulations of each variable by FCTR to ascertain how to cut the variable and still maintain maximum discriminating power with respect to FCTR. The following dichotomizations were made:

	"low" values	"high" values
YEARS	0 - 3	4 and over
NEWS	"never" and "occasionally"	"Often"
HLANG	"no response" and any response except "English"	"English"
GRADE	0 through 6th grade	7th grade and above

The basic crosstabulation, then, was SPUND x YEARS x NEWS x HLANG x GRADE. It had a total of 144 cells. Each cell represented a particular pattern of MELP responses, and within each cell was placed the number of adults displaying that response pattern and the proportion of those who were classified LESA by FCTR. That crosstabulation is reproduced as Table 10. (In it SPUND categories 2-7 have been collapsed to facilitate presentation.)

Table 10: Percent LESA adults for various combinations of SPUND, YEARS, NEWS, HLANG, and GRADE. ND indicates no respondents in that cell. ( ) indicates N in cell.

NEWS=low
HLANG=low
GRADE=low

SPUND	YEARS Low	YEARS High
2-7	72 (25)	0 (1)
8	67 (9)	20 (5)
9	ND	0 (2)
10	ND	0 (1)

NEWS=high
HLANG=low
GRADE=low

SPUND	YEARS Low	YEARS High
2-7	92 (165)	64 (11)
8	44 (9)	25 (4)
9	0 (1)	100 (1)
10	100 (1)	0 (2)

NEWS=low
HLANG=high
GRADE=low

SPUND	YEARS Low	YEARS High
2-7	100 (6)	100 (1)
8	50 (2)	0 (2)
9	0 (1)	ND
10	ND	0 (1)

NEWS=high
HLANG=high
GRADE=low

SPUND	YEARS Low	YEARS High
2-7	63 (8)	ND
8	0 (2)	0 (3)
9	100 (2)	ND
10	ND	ND

NEWS=low
HLANG=low
GRADE=high

SPUND	YEARS Low	YEARS High
2-7	74 (27)	11 (19)
8	24 (21)	13 (23)
9	20 (5)	0 (11)
10	ND	0 (25)

NEWS=high
HLANG=low
GRADE=high

SPUND	YEARS Low	YEARS High
2-7	76 (156)	38 (34)
8	36 (42)	17 (41)
9	0 (2)	14 (7)
10	0 (1)	0 (11)

Table 10 continued

NEWS=low
HLANG=high
GRADE=high

SPUND	YEARS Low		YEARS High	
2-7	100	(2)	0	(5)
8	50	(4)	03	(29)
9	ND		0	(6)
10	33	(3)	03	(73)

NEWS=high
HLANG=high
GRADE=high

SPUND	YEARS Low		YEARS High	
2-7	100	(3)	0	(2)
8	0	(3)	06	(18)
9	ND		20	(10)
10	ND		0	(18)

The Cochran and Hopkins procedure calls for assigning to the category LESA any cell which has a larger proportion of LESAs than does the sample as a whole. In this case, 49% of the total sample of adults are LESAs, so any cell with 50% or more LESAs was considered to be LESA. Using this procedure on the entire 144 cell table, the following table was derived representing the predictive accuracy of the five variables relative to dichotomized FCTR.

		FCTR LESA	FCTR -LESA	Total
Predicted	LESA	86%	16%	
		384	76	460
	-LESA	14%	84%	
		60	395	455
	Total	444	471	915

85% categorized the same
% LESA by FCTR= 49
% LESA by MELP= 50
% Bias= 4

This represents the maximum correspondence that any explicit operational definitions of LESA and non-LESA involving these five variables could have with FCTR.

An examination of Table 10 indicates that the most powerful predictor variables were SPUND and YEARS. Performing a Cochran and Hopkins analysis on just these two predictors, the following table was derived:

Predicted (SPUND, YEARS)	FCTR LESA	FCTR -LESA	Total
LESA	82%	17%	
	363	78	441
-LESA	18%	83%	
	81	393	474
Total	444	471	915

83% classified the same
49% classified LESA by FCTR
48% classified LESA by MELP
-1% Bias

The pattern of cells underlying the above table happen to exactly conform to the following definitions of LESA and non-LESA.

1. A respondent is non-LESA if: (SPUND > 8) or(YEARS > 3)
2. A respondent is LESA if he has any other values of SPUND and YEARS.

The reduction of five predictors to two predictors loses only two percent in the number of respondents classified the same by MELP and FCTR, and the amount of bias remains very low for the sample of adults as a whole. Thus, it is this definition that we would choose for adults. The accuracy of the definition within each ethnic group is given in Table 11 and 12. Percent categorized the same by the definition and FCTR ranged from 76 to 90. The absolute difference between percent identified as LESA and FCTR and that identified as LESA by the definition varied from essentially zero to 8% while the percent bias varied from a 6% over-estimation to 53% underestimation.

Table 11: Performance of SPUND-YEARS scoring key by group.

A. Cubans

		FCTR		
	LESA	-LESA	Total
LESA	86%	44%	
	159	38	197
MELP			
-LESA	14%	56%	
	26	49	75
Total	185	87	272

B. Chicanos

		FCTR	
	LESA	-LESA	Total
LESA	91%	50%	
	146	21	167
MELP			
-LESA	9%	50%	
	14	21	35
Total	160	42	202

C.

		FCTR	
	LESA	-LESA	Total
LESA	80%	18%	
	45	10	55
MELP			
-LESA	20%	82%	
	11	45	56
Total	56	55	111

D.

		FCTR	
	LESA	-LESA	Total
LESA	24%	4%	
	4	4	8
MELP			
-LESA	76%	96%	
	13	95	108
Total	17	99	116

E. Navajo

		FCTR	
	LESA	-LESA	Total
LESA	35%	3%	
	9	5	14
MELP			
-LESA	65%	97%	
	17	183	200
Total	26	188	214

Table 12: Accuracy of SPUND-YEARS scoring key by group.

	Overall	Cuban	Chicano	Chinese	Other Asian	Navajo
% categorized the same by FCTR and MELP	83	76	83	81	85	90
% categorized LESA by FCTR	49	68	59	50	15	12
% categorized LESA by MELP	48	72	61	50	7	7
% Bias	-1	+6	+3	0	-53	-42

5. Recommended Scoring Keys for Categorizing Adults as LESA and non-LESA

On the basis of the analysis detailed above, two alternative scoring keys are recommended for categorizing adults as LESA and non-LESA on the basis of their MELP responses:

A. A discriminant function involving eleven predictor variables. The discriminant function was derived by pooling all adult data into a single analysis. The equation is as follows:

$$Y_{FCTR} = -.08*WHEN - .25*SPEAK - .13*UNDERSTAND - .05*KID - .05*FRIEND + .12HLANG - .05*YEARS + .19*NEWS - .10*BIRTH - .03*GRADE - .06*INCOME + 2.06$$

For any given respondent, if his discriminant function score is above 0.02 he is assigned to the LESA category. If his score is equal to or below that value, he is assigned to the non-LESA category.

B. An operational explicit definition involving the variables SPUND and YEARS. An adult is assigned to the category non-LESA if his response pattern conforms to either of the following patterns:

1. SPUND greater than 7
2. YEARS greater than 3

All other adults are assigned to the LESA category.

With respect to overall performance, these two scoring keys are approximately equivalent; however, they were derived using markedly different approaches. The discriminant function approach is basically a multiple regression approach and its strengths and weaknesses are well-known. For example, it assumes continuous predictors (which we clearly do not have). The contingency table approach requires very few assumptions; however, the data from the field test are relatively sparse in some regions of the table and thus generalizing from them may be risky. Which scoring key is used depends on an individual's preference.

IX. Finding an Unbiased Estimator of the Proportion of LESAs in the U. S.

In Chapters VII and VIII, we have developed scoring keys which give relatively useful results for predicting the dichotomized LESA distributions of the respondents in our field test; however, the sampling plan of the field test differed from that of the SIE in important ways and the ramifications of these differences must now be considered. Many of the issues raised in this chapter and the solutions proposed to deal with them were spelled out in a conference attended by representatives of CAL, RTI, and NCES and by a number of our technical consultants. The proceedings of that conference and the list of participants can be found in Appendix 17.

1. The List Samples

With respect to children, using list samples delivered to us by the schools had several advantages. The sampling required almost no statistical expertise or prior knowledge of the communities on the part of RTI or CAL. Also, the dichotomous property of the lists was invaluable for constructing scoring keys that yielded dichotomous classifications. However, the use of lists also had disadvantages. The first disadvantage was that RTI and CAL essentially lost control of how children were selected onto the lists from the pool of all children in the school districts who had been screened for their English proficiency. Thus, we have no grounds for assuming that the lists in the various sites were random samples of all the children in that age range who were so classified or that the interviews obtained were a random sample of the lists obtained from the schools. For example, in each site, about a third of the addresses given as the children's residences were found to be wrong, and informal evidence indicated that some parents deliberately gave the school incorrect addresses to avoid busing or some other administrative regulation. In the majority of cases such children were simply replaced with

others from the same list. Also, in San Francisco, lists were constructed from the rosters of only a few schools selected for their high concentrations of the ethnic groups we were examining. These are just two of the factors that caused the samples of children interviewed to be decidedly non-random parts of the LEA's potentially LESA populations.

The sampling problem with adults was also serious. In Miami and El Paso, all adults were sampled from the pool of individuals who had recently enrolled in adult education classes. How that pool relates to the general pool of non-native-English speakers in those areas as legislatively defined is completely unknown. In Arizona and San Francisco, where adults were taken from the households of the children, all of the sampling problems of the children apply to the adults with the additional qualification that all these adults came from households containing an elementary school aged child.

2. The Distribution of LESAs and non-LESAs

A second, quite different problem was that RTI was instructed to interview approximately equal numbers of individuals on each of the lists they obtained. This led to the production of scoring keys which had approximately equal error rates for the identification of LESAs and non-LESAs. However, we have reason to suppose that the two categories are not at all in equal proportions nationally. A recent census of the Spanish speaking school population of Dade County (Florida) indicates children on the "independent" list to be three to four times more numerous than the children on the other two lists combined. Similarly, but more indirectly, preliminary tabulations from the July, 1975 "Survey of Languages," done by the Bureau of the Census for NCES, indicates that a large majority of school children whose native language (as defined legislatively) is not English are reported by Household Respondents to have "no difficulty" in speaking or understanding English.

The problem is that if the scoring key is to provide an unbiased estimate of the proportion of LESA children, its rates of identification errors must be proportional to the relative numbers of LESA and non-LESA children in the population. To illustrate: Suppose we have a procedure which identifies both LESA and non-LESA children with an error rate of 25%. This could be expressed in a four fold table as:

		True Category	
		LESA	non-LESA
Application of fallible procedure	LESA	75%	25%
	Non-LESA	25%	75%
		100%	100%

If this fallible procedure were applied to a population with equal numbers of LESAs and non-LESAs, it would yield an unbiased estimate of the population proportion of LESAs since it would falsely identify the same numbers of LESAs and non-LESAs. Applied to a population of 1000:

		True Category		Estimated Totals
		LESA	non-LESA	
Application of MELP Procedures	LESA	75%	25%	
		375	125	500
	non-LESA	25%	75%	
		125	375	500
Actual Totals		500	500	
		(100%)	(100%)	

75% categorized correctly by MELP
50% True LESAs
50% Categorized LESA by MELP
0% Bias

However, now consider the same procedure applied to a population of 1000 where the true number of LESAs is only 200:

		True Category		Estimated Totals
		LESA	non-LESA	
Application of Procedure	LESA	75%	25%	
		150	200	350
	non-LESA	25%	75%	
		50	600	650
Actual Totals		200	800	

75% Categorized correctly by MELP
20% True LESAs
35% Categorized LESA by MELP
+75% Bias

In this case, while the procedure still errs at the same rate in each category, the resulting estimate of the true proportion of LESAs is highly biased -- 35% as compared with a true proportion of 20%. Clearly what is needed is a revised procedure which will mis-classify equal numbers of children rather than equal percentages. But this involves adjusting the error rates in a ratio equal to the ratio of LESAs to non-LESAs in the population.

For example, if the true ratio of LESA to non-LESA persons in the population was one to four, as in the above example, then in order to misclassify equal numbers of individuals the procedure would have to identify non-LESAs with an error rate one fourth the magnitude of the error rate involved in identifying LESAs. This could yield the following table:

		True Category: LESA	True Category: non-LESA	Estimated Totals
Application of Procedure	LESA	37%	16%	
		75	125	200
	non-LESA	63%	84%	
		125	675	800
Actual Totals		200	800	

75% Categorized correctly by MELP
20% True LESAs
20% Categorized LESA by MELP
0% Bias

We have chosen the numbers in this table so that it has the same total number of individuals categorized correctly as the table above it (75%) -- in other words, the two procedures have the same overall validity. However, in this latter case, the procedure does very badly in identifying LESAs (classifying more wrong than

right), and exactly four times better (63%/4=16%) in identifying non-LESAs. This leads to balanced numbers of false positive LESA identifications and false negative ones. There are a number of ways in which empirically-derived identification procedures such as those in Chapters VII and VIII can be calibrated to display a particular ratio of false positive and false negative identifications; but in order to do such a calibration, either the true proportion of LESA individuals in the population must be estimated in advance (to estimate this is the reason for the survey in the first place) or the "true" error rates of indentification in the population of interest the SIE population must be known. Unfortunately, because of the sampling factors already discussed, we can have no confidence in estimating these from the field test results.

This problem is treated in depth both from theoretical and empirical perspectives by Hartwell et. al. in the Research Triangle Institute's final report to CAL on their subcontract for this project. The reader is referred to section V.F.4 and page 100 in that report. (Hartwell, Moore, Weeks, Mason, and Shah; Design, Data Collection and Analysis of Instruments and Procedures to Measure English Language Proficiency. Research Triangle Park, North Carolina: Research Triangle Institute. April, 1976.)

Basically, Hartwell explored three ways of coping with the problem. First, he artifically simulated the expected relative proportions of LESA and non-LESA respondents in the nation by creating a new data file in which all non-LESA data in the field test corpus was duplicated 4 times. This new file was then subjected to discriminant analysis. The results indicated that the discriminant functions derived from the new file were similar to the original functions, but that they over-estimated the percent of LESA individuals in all groups by 8% to 144%. Overall, the overestimation was 28%.

Second, Hartwell explored the use of a correction factor that could be applied to the SIE data to estimate the percent LESA children nationally. This correction factor, however, assumes that the user has accurate estimates of the rates at which the identification procedure (the MELP) produces false positive and false negative identifications of LESA in the SIE context. The only estimates available are from the field test data, and they are suspect because of the non-random selection of respondents within the LESA and non-LESA categories already discussed. Nevertheless Hartwell, et. al., present evidence that such a post hoc correction may be more accurate than attempting to derive a usable scoring key through simulation techniques. This procedure generally produced under-estimates of LESA proportions (in four of the five groups) rather than the over-estimates resulting from the simulations. The deviations of the estimations from percent LESA as defined by list ranged from an overestimation of 30% in % bias terms in the Other Asian group to an underestimation of 39% in the Cuban group.

A third technique, favored over the other two by RTI, was for a two-stage sampling plan to be executed as part of the SIE. This would involve obtaining criterion information -- perhaps both list and FCTR or DORP -- on a representative subsample of children from the SIE households as soon after their regular interview data were gathered as possible. From this information, accurate national estimates of the percent of LESA could be derived and the scoring keys could either be rederived or recalibrated.

3. Adjusting the Face-Valid Definitions of LESA and non-LESA

In pursuing the recalibration of the MELP to accomodate it to the expected low proportion of LESAs in the SIE sample, RTI only worked with the discriminant analysis, however, the face-valid definitions can also be adjusted to give a more accurate estimate of LESAs in the SIE context. Basically what desired is to

find a definition which, when applied to the field test data, will yield a ratio of false positives of LESA identification to false negatives equal to the ratio of LESAs to non-LESAs in the SIE sample. In particular,

$$\frac{\text{proportion of False Negatives}}{\text{proportion False Positives}} = \frac{\text{proportion of non-LESAs}}{\text{proportion of LESAs}}$$

Let us assume that for children the above ratio is four to one in the general population of non-native English speaking children and what is required is a modification of Definition 2 to accommodate it to this fact. Over the entire field test sample, Definition 2 yielded a ratio of 99/133 or .74. What is needed is to redefine the non-LESA category to include more respondents. Applying the Cochran and Hopkins procedure to this situation, p becomes .80 as the criterion for deciding whether a given SPUND-USE-YEARS response pattern is to be considered LESA or not. Table 19 in Section VII has 13 cells with percent LESA above 80. A possible definition might be:

Definition 3: A child is non-LESA if his response pattern meets either of the following conditions:

1. A USE score of 2 or 3
2. A SPUND score of 8,9, or 10

The correspondence of this definition with list is:

		LIST	
	LESA	non-LESA	Total
LESA	50%	7%	
	323	33	356
Definition 3			
non-LESA	50%	93%	
	319	423	742
Total	642	456	

60% classified the same by List and Def. 3
58% classified LESA by List
32% classified LESA by Definition 3
-45% bias.

Now, suppose we artifically simulate a "true" LESA - non-LESA ratio of 1 to 4 from the field test data by simply multiplying the list non-LESA column by 5.63, obtaining:

		LIST	
	LESA	non-LESA	Total
LESA	50%	7%	
	323	186	509
non-LESA	50%	93%	
	319	2382	2701
Total	642	2568	3210

84% classified the same by List and Def. 3
20% classified LESA by List
16% classified LESA by Def. 3
-21% Bias

Definition 3 in a sense overshoots its objective by classifying too few respondents as LESA even with only 20% actual LESAs in the population. (This compares with an overestimation of 101% if Definition 2 were applied to the above simulated population.) Clearly, Definition 3 could be modified slightly to categorize a slightly larger % of respondents as being LESA. (For example, changing condition 1 of Definition 3 to include USE values of only 3 would result in 29% of the simulated population being categorized as LESA.) Such "fine tuning" of these definitions has a completely ad hoc character with unknown generalizability beyond the samples involved here. Nevertheless, it is important to note that a number of reasonably face-valid definitions can be easily formulated, each with distinct implications for the magnitudes of the LESA counts obtained through their use.

4. Summary of Recalibration Recommendations

CAL recommends RTI's double sampling proposal with the added recommendation that the face-valid definitions suggested above and some similar ones be tested on the data obtained in the double sampling effort. This would be in addition to re-deriving the disciminant functions using those data. If such a double sampling is not possible to implement, then the correction formula suggested by Hartwell can be used. (Note, however, that Hartwell's cautions on page 94 are extremely important). In any case, the behavior of a scoring key in estimating different proportions of LESA individuals must be kept in mind. That is, if

$$\frac{P\ (\text{False Negatives})}{P\ (\text{False Positives})} = \frac{P\ (\text{non-LESAs in the population})}{P\ (\text{LESAs in the populations})}$$

then the number of LESAs in the population will be systematically under-estimated while if the inverse obtains the number of LESAs will be over-estimated. Putting

it slightly differently, if we estimate the ratio of error rates for some scoring key to be, say 50%/7% or 7.14 (as for Definition 3), then we know that for populations in which the true ratio of non-LESAs to LESAs is less than that, there will be an underestimation of LESAs. Thus, if Definition 3 were applied to the SIE data and yielded a % LESA of 30, we would know that that was an underestimate. (Again, this assumes that the error rates as observed in the field test are reasonably accurate.) On the other hand, if we obtained an estimate of 30% using Definition 2, then we would know that it was an overestimate since the ratio of error rates for Definition 2 is 0.52 while the observed ratio of non-LESAs to LESAs was 70%/30% or 2.33.

5. Adults

The proposal for double sampling and recalibration of the scoring keys applies only to children because it is only for them that there are relatively unequivocal dichotomous classifications of LESA and non-LESA available externally to the SIE (that is, from schools). If adults were to be double-sampled, the criterion instruments which could be used would be a test or a direct rating. Neither of these, however, has a non-arbitrary way of dichotomizing the scores obtained from them into LESA and non-LESA categorizations. Thus, the criterion instruments would not lead to a robust estimation of the proportion of LESA adults in the nation.

The alternatives for adults would seem to be two:

1. Use the discriminant function to estimate LESAs and simply keep in mind that if the obtained proportion deviates greatly from .5 (approximately what it was in the field test), it is a biased estimate; i.e., P(LESA) .5 implies a probably underestimation and P(LESA) < .5 implies a probable overestimation.

2. Use the face-valid definition and simply depend on its manifest content to provide an accurate count of LESAs. This amounts to assuming that if an adult is claimed to have spent more than three years in an English-language school or is claimed to speak and understand English well, then he is counted as non-LESA.

X. Accuracy of MELP Data as Reported by a Household Respondent about Another Adult in the Household.

In the SIE, the Household Respondent, will generally be the source of all information about each individual in the household. The purpose of this section is to explore the quality of the data given by the Household Respondent about another adult (14 years old and older) member of the household. Such data, which we will call proxy data, will be examined both for its correspondence with first-hand-data -- information which an adult gives about himself -- and for its correspondence with dichotomized FCTR. It should be noted that this problem arises only with adults because children (13 and younger) will never be asked to give information to the SIE interviewer; all information about the children in a household will be obtained from the Adult Household informant. Thus, all child data will be proxy data.

So far in this report, all analyses that have been reported for adults have been based on first-hand data -- that an individual gave about himself. (All child data analyses in this report are based entirely on proxy data.) However, during the field test, whenever there was an adult available in the household in addition to the Designated Adult Respondent, he or she was asked to serve as a Household Respondent and provide answers to the MELP questions about the Designated Respondent. Unfortunately, in many households there was not an appropriate additional person available so proxy data were unobtainable.

The following table indicates the amounts of proxy data available in the various groups for analysis:

Group	Adult Respondents	Household (proxy) Respondents	Proportion proxy data
Cuban	272	118	.43
Chicano	202	96	.48
Chinese	111	45	.41
Other Asian	116	48	.41
Navajo	214	178	.83
Total	915	485	.53

A first question to be asked is whether the proxy respondent gives answers at all to the MELP questions. Table 1 gives the percent of answers of "don't know" or "no answer" for proxy data for each MELP question.

Table 1: Percent Scoreable and Unscoreable Answers given by Household Respondents about Other Adult Members of the Household. (All Groups Combined, N= 485)

MELP Variable	Scoreable Response	Don't Know	No Answer
WHEN	96	4	0
SPEAK	97	0	3
UNDERSTAND	97	0	3
SIB	100	0	0
FRIEND	99	1	0
HLANG	99	0	1
YEARS	84	3	13
BIRTH	84	11	5
GRADE	92	7	1

It clearly shows that the rate of usable responses is very high for all variables except the "historical" ones -- that is, those asking for specific facts about an individual's background -- in which case 8 to 16% of the responses were either not recorded or "don't know". These rates may be higher than those to be encountered in the SIE proper for the following reason: In the field test, the household respondent was instructed to answer the questions about the designated respondent on the basis of his <u>own</u> knowledge. There was to be no "pooling" of information from any and all members of the household present at the time. In the SIE, however, the interviewer will make an effort to obtain complete information on each household member from whomever is available at the moment. In other words, the interviewer will not be compelled to talk to only one individual per household. Thus, we might expect more complete information using that procedure.

An important statistic to be derived from the field test data is the number of usable protocols that could be entered into a scoring key and thus from which a LESA - non-LESA categorization could be derived. In the case of the first-hand data, the total sample for which FCTR scores were available was 1150 while the total number for which there was complete MELP data was 915 or approximately 80%. In the case of the proxy data, there were 454 FCTR scores and 313 complete MELP protocols (69%). Thus, if these data provide reasonable guidance, NCES should expect up to 10% fewer complete protocols derived from proxy data than from first-hand data.

A second question about proxy data is: are the data obtained as predictive of LESA and non-LESA categorizations as are first-hand data? To answer this question, the overall discriminant function was applied to the 313 proxy protocols. The following table indicates the resulting correspondence with dichotomized FCTR scores:

	FCTR LESA	non-LESA	Total
LESA			
Discr. Function MELP (using proxy data)	118	38	156
non-LESA	19	138	157
TOTAL	137	176	313

82% categorized the same by Test and MELP
44% categorized LESA by test
50% categorized LESA by MELP
+12% Bias

These figures are highly similar to those for first-hand data. There the percent categorized the same by test and MELP was 83% and the bias was + 11%. The correspondence with test of the SPUND-YEARS definition of LESA non-LESA when applied to proxy data are given below:

	TEST LESA	non-LESA	Total
LESA			
Definitional MELP (proxy data)	107	25	132
non-LESA	30	151	181
TOTAL	137	176	313

82% categorized the same by test and MELP
44% categorized LESA by test
42% categorized LESA by MELP
- 4% Bias

Again these figures are highly similar to those for first-hand data.

Finally, Table 2 gives the cross tabulations of the first-hand and proxy data for the three most important MELP variables: SPEAK, UNDERSTAND, and YEARS.

Table 2: Cross tabulations of Proxy by First Hand Responses to Selected MELP Questions

A. SPEAK

		Proxy Response					
		1	2	3	4	5	Total
	1	34	16	1	4	1	56
	2	16	133	11	27	6	193
First Hand	3	1	5	1	6	0	13
Response	4	0	12	3	70	41	126
	5	1	3	0	17	76	97
Total		52	169	16	124	124	485

Response options:

1 = not at all

2 = Just a little, don't know, or missing data

3 = Adequate for a few purposes

4 = Well, adequate, or adequate for most purposes

5 = Very well

B. UNDERSTAND — Proxy Response

First Hand	1	2	3	4	5	Total
1	31	9	3	3	0	46
2	13	102	11	30	4	160
3	1	7	5	11	2	26
4	1	17	4	84	55	161
5	1	5	1	12	73	92
Total	47	140	24	140	134	485

C. YEARS — Proxy Response

First Hand Response	0	1	2	3	4	5	>5	Missing or DK	Total
0	90	8	3	3	1	0	1	16	122
1	7	32	2	3	0	0	0	7	51
2	8	6	11	1	2	0	1	10	39
3	1	0	5	15	2	0	4	7	34
4	1	1	1	2	3	1	2	1	12
5	0	0	0	2	0	5	4	4	15
5	2	1	2	1	1	2	163	17	189
Missing or DK	5	2	1	0	0	0	0	15	23
Total	114	50	25	27	9	8	175	77	485

It is clear from this table that proxy responses are generally similar to first-hand responses. In Tables 2A and 2B, 65% and 61% of the responses respectively were identical for first-hand and proxy respondents. In Table 2C, 66% of the

responses fall on the main diagonal. For both SPEAK and UNDERSTAND, there is a tendency for the Household Respondent to rate the Designated Respondent slightly higher than the Designated Respondent would rate himself. The mean first-hand rating for SPEAK is 3.03 while the mean proxy rating is 3.20. For UNDERSTAND the mean first-hand rating is 3.19 compared to 3.36 for the mean proxy rating. (These differences are both statistically reliable at $p < .01$). Such a tendency is not apparent in the YEARS variable, with means of 5.18 and 5.16 for the first-hand and proxy responses respectively. In fact, there were no statistically reliable tendencies for first-hand and proxy data to differ systematically from each other on any of the other MELP variables. Given the slightly higher ratings in proxy data, one would expect a correspondingly slight tendency for estimating fewer LESAs from proxy data than from first hand data. Assuming the discriminant function to be roughly normally distributed that difference would be about 2%.

Summary

Comparisons of the data elicited from Household Respondents (proxy data) and Designated Respondents (first hand data) lead to the following generalizations:

1. On questions calling for specific information about a person's background (birth date, education, etc.), there were approximately 10% fewer responses given by Household Respondents than by Designated Respondents. Different interviewer instructions in the SIE should result in a smaller percentage of "Don't know" and "No Response" codes being transcribed.

2. On all other MELP questions there was essentially complete data from Household Respondents.

3. The overall relationship of proxy data to FCTR was very similar to that of first-hand data.

4. On SPEAK and UNDERSTAND there were slight but significant tendencies for proxy ratings to be higher than first-hand ratings. This could lead to an underestimation of LESAs from proxy data of about 2% relative to first-hand data.

XI. A Comparison of Monolingual and Bilingual Interviewers

In the SIE, most of the interviewers will be able to speak and understand only English. Concern was expressed by both LGRs and technical consultants that this type of interviewer might be less effective in obtaining accurate information from a potential LESA individual than a bilingual interviewer (i.e., an individual who speaks English and the language of the respondent). As a result of this concern, a study was conducted during the field test to compare the effects of monolingual vs. bilingual interviewers on data collection results.

Over all sites 101 interviewers were employed: 50 were monolingual (English) and 51 were bilingual, speaking both English and the language of the respondent, and were members of the respondents' ethnic groups. Within each site five pairs of interviewers were assigned to work in five separate sub-areas of the site. Each pair consisted of one monolingual and one bilingual interviewer. The interviewers were randomly selected to participate in the substudy and the sample cases assigned to each pair member were randomized. In the San Francisco site, only Chinese bilingual interviewers were available. Thus, Other Asians did not participate in this study.

Instructions to the interviewers for administering the census type questions (including the MELP questions) were as follows: All interviewing was to be carried out in English whenever possible. If communication with the respondent was too difficult or inaccurate, then one of two courses of action was to be taken:

1. Bilingual interviewers were to switch to the respondent's native language whenever necessary.

2. Monolingual (English-speaking) interviewers were to find another individual, either in the household or from the neighborhood, who could act as translator.

Of course, both the tests and the DORP were administered entirely in English.

Three different analyses were run on the data to compare bilingual vs. monolingual effects. Each of these will be described below.

1. Production Data

One way in which bilingual and monolingual interviewers could differ is in the number of interviews completed. In fact, several LGRs predicted in June that monolingual Anglo interviewers would be faced with a much higher rate of refusals to be interviewed than would interviewers who were members of the respondents' ethnic-linguistic group. Thus, the expectation was that with respect to gross quantity of data collected, bilingual interviewers would be more productive than monolingual interviewers. (It should be noted that monolinguals' instructions were to find someone in the neighborhood to translate if communication with all members of the household was insufficient to conduct the interview. Bilingual interviewers were instructed to conduct the interview in English whenever possible and to use the other language only when absolutely necessary.)

Table 1, reproduced from RTI's final report (their table IV.4) summarizes various production statistics for monolingual and bilingual interviewers. Notice that this is for all interviewers, not just the ten in each location who were matched with each other and it is for all respondents, both child and adult. There appear to be no large differences between the two types of interviewers in terms of the number of respondents interviewed. In fact, the monolingual interviewers completed more (64%) interviews than did the bilingual interviewers (61%). This discrepency is statistically significant for the Navajos and Chinese and also for all groups pooled together ($2 = 3.49$, 2.16 and 2.11, respectively).

Refusal rates were low in all groups, and there were no significant differences between monolingual and bilingual interviewers in this regard. These results do

Table 1

COMPARISON OF DATA COLLECTION RESULTS FOR MONOLINGUAL AND BILINGUAL INTERVIEWERS[1/]

	Miami		El Paso		Arizona		San Francisco		Total	
	Mono.	Bi.	Mono.	Bi.	Mono.	Bi.	Mono.	Bi.	Mono.	Bi.
No. of Interviewers[2/]	10	15	9	14	10	13	21	9	50	51
Potential Respondents Assigned[3/]	404	675	397	674	533	439	803	389	2137	2177
Respondents Interviewed	249	419	260	431	394	279	470	202	1373	1331
(Percent)	(62%)	(62%)	(65%)	(64%)	(74%)	(64%)	(59%)	(52%)	(64%)	(61%)
Refused	11	15	7	11	8	8	39	15	65	49
(Percent)	(3%)	(2%)	(2%)	(2%)	(2%)	(2%)	(5%)	(4%)	(3%)	(2%)
Other Nonrespondents[4/]	144	241	130	232	131	152	294	172	699	797
(Percent)	(36%)	(36%)	(33%)	(34%)	(25%)	(35%)	(37%)	(44%)	(33%)	(37%)
Total Nonrespondents	155	256	137	243	139	160	333	187	764	846
(Percent)	(38%)	(38%)	(35%)	(36%)	(26%)	(36%)	(41%)	(48%)	(36%)	(39%)
Total Hours Charged[5/]	1099	1817	1182	1810	1717	1486	1937	980	5935	6093
Total Miles Driven[6/]	9412	13554	8106	12973	18929	15399	7088	1211	43535	43137
Average Hours Per Interview	4.4	4.3	4.6	4.2	4.4	5.3	4.1	4.9	4.3	4.6
Average Miles Per Interview	37.8	32.4	31.2	30.1	48.0	55.2	15.1	6.0	31.7	32.4

1/ Figures in this table are based upon manual counts and computations by interviewers and supervisors and have not been verified by machine tabulations.

2/ All interviewers spoke English. For purposes of this study, "monolingual" referred to interviewers who did not also speak the language of the respondent, while "bilingual" interviewers did speak the respondent's language.

3/ In Miami and El Paso both children and adults were assigned to interviewers. In Arizona and San Francisco only children were assigned, since no adult lists were obtained for these sites. Interviewers randomly selected an adult from each sample child's household in these sites. For Arizona and San Francisco, therefore, the number of potential respondents was twice the number of sample children assigned.

4/ Examples of "other" nonrespondents include cases where the sample member had moved to another city; where the address was nonexistent; where the sample member could not be contacted at home in the prescribed number of interviewer visits; where the sample member was out of town; or where he was sick, institutionalized, or otherwise unavailable.

5/ Includes training time.

6/ Includes mileage incurred in connection with training.

not support the predictions of the LGRs that monolingual interviewers would have difficulty obtaining interviews. One factor that may have played a role here is that the monolinguals, as a group, had more prior experience in interviewing than did the bilinguals. Only 11 of the 51 bilinguals had interviewing experience prior to this project while 22 of the 50 monolinguals were experienced interviewers.

2. Comparisons of MELP and Test Data as Gathered by Monolinguals and Bilinguals.

A second way in which monolingual and bilingual interviewers could differ was in the quality of the data they collected. In other words, were the responses to some MELP questions and/or test items biased by the language ability and ethnic group membership of the interviewer? To answer this question, the means of the various MELP variables and the test total scores were compared for the matched data of the five pairs of interviewers in each site.

Child Data. Table 2, after Table V.32 of RTI's final report, gives the means for children and the results of t-tests on them. Out of 55 comparisons, there were only three that were significantly different. Two of these occurred in El Paso:

Table 2: (Reproduced from RTI's Report, Table V.32) Sample Means and Summary of t-tests on Monolingual Versus Bilingual Interviewer Means for Various MELP questions and Test Total Score, Data for Children for Paired Interviewers Only.

Variable	Interviewer Type	Cubans	Chicanos	Navajos	Chinese	Over Groups
When	Mono	1.90 ns	2.75 ns	2.98 ns	2.24 ns	2.49 ns
	Biling	1.76	2.85	3.00	2.58	2.50
Speak	Mono	3.38 ns	3.22 *	3.78 ns	3.40 ns	3.46 ns
	Biling	3.23	3.76	3.66	3.65	3.60
Und	Mono	3.66 ns	3.45 ns	4.00 ns	3.72 ns	3.71 ns
	Biling	3.43	3.69	3.89	3.69	3.68
Sib	Mono	1.65	1.95	1.78	2.04	1.83
	Biling	1.65 ns	1.76 ns	1.86 ns	2.15 ns	1.84 ns
Frnd	Mono	2.16 ns	2.23 *	2.13 ns	2.56 ns	2.23 ns
	Biling	2.02	1.87	2.23	2.46	2.14
Hlang	Mono	1.03 ns	1.77 ns	1.83 ns	1.52 ns	1.53 ns
	Biling	1.00	1.75	1.77	1.46	1.51
Years	Mono	2.47 ns	1.70 ns	3.34 ns	2.84 ns	2.58 ns
	Biling	2.22	1.85	3.93	2.35	2.54
Birth	Mono	65.9 ns	67.5 ns	65.6 ns	67.2 ns	66.4 ns
	Biling	65.9	67.5	65.2	67.0	66.4
Grade	Mono	5.03 ns	3.77 ns	5.27 ns	4.16 ns	4.65 *
	Biling	4.90	3.02	5.30	3.85	4.22
Ped	Mono	2.85 ns	2.87 ns	2.89 ns	3.68 ns	2.97 ns
	Biling	2.71	3.00	2.75	3.73	3.08
Test	Mono	44.1	41.9	50.2	46.6	45.6
	Biling	41.6 ns	39.3 ns	50.1 ns	49.8 ns	44.4 ns
Sample	Mono	68	60	64	25	220
	Biling	51	55	44	26	186

* = t-Test significant at .05 level. ns = t-Test not significant at .05 level.

(1) In response to the question "how well does . . . . speak English," respondents tended to give higher assessments when asked by a bilingual vs. a monolingual interviewer.

(2) In response to the question "what language does . . . . speak to his friends", El Paso respondents claimed "English" slightly more often to monolingual than to bilingual interviewers. Only one overall comparison was significant: monolingual interviewers were told that their respondents were in a slightly higher grade than were bilingual interviewers. This is evidenced by the "Grade" comparison. The interpretation of this finding is relatively unclear for several reasons:

1. Since overall completed interviews averaged less than two-thirds of total assignments, the random assignment of interview loads to the members of each pair may not have been preserved in the completed interviews. Thus, it is possible that monolingual interviewers had a slight tendency not to complete interviews with children in the lower grades. However, it could also be that parents merely tend to report a higher grade to monolingual interviewers.
2. One would expect that higher values of GRADE would be accompanied by different BIRTH values, but such was not the case.
3. The tendency was not replicated across groups in a consistent way.

Finally, it should be noted that there were no mean test score differences between those tests administered by monolingual and bilingual interviewers.

Adult Data. A similar comparison of means is presented in Table 3 for adult (first-hand) data. In this case, across all groups, monolingual-interviewed respondents scored significantly higher on the test than did bilingual-interviewed respondents. They also scored significantly higher on the SPEAK, UNDERSTAND, and INCOME variables.

Table 3: Sample Means and Summary of t-Tests on Monolingual Versus Bilingual Interviewer Means for Various MELP Questions and the Total Test Score, Data for Adults and Paired Interviewers Only

Variable	Interviewer Type	Group: Cubans	Chicanos	Navajos	Chinese	Over Groups
When	Mono	1.55	1.97	-	1.82	2.19
	Biling	1.60^{ns}	2.14^{ns}	-	1.95^{ns}	2.09^{ns}
Speak	Mono	2.57	2.16	4.22	3.00	3.16
	Biling	2.19^{ns}	2.51^{ns}	3.71*	2.90^{ns}	2.77**
Under-stand	Mono	2.98	2.18	4.23	3.06	3.29
	Biling	2.43*	2.67^{ns}	3.88^{ns}	2.95^{ns}	2.94*
Kid	Mono	1.47	1.24	2.10	1.65	1.66
	Biling	1.51^{ns}	1.53^{ns}	1.88^{ns}	1.65^{ns}	1.64^{ns}
Friend	Mono	1.14	1.11	2.11	1.53	1.52
	Biling	1.19^{ns}	1.35^{ns}	1.88^{ns}	1.65^{ns}	1.46^{ns}
Hlang	Mono	1.06	1.18	2.05	1.47	1.47
	Biling	1.02^{ns}	1.33^{ns}	1.75^{ns}	1.25^{ns}	1.34^{ns}
Years	Mono	1.61	1.71	9.16	5.06	4.83
	Biling	0.77*	1.60^{ns}	8.97^{ns}	5.20^{ns}	3.81^{ns}
News	Mono	1.98	2.08	1.56	2.24	1.87
	Biling	2.15^{ns}	2.02^{ns}	1.78^{ns}	2.20^{ns}	2.03^{ns}
Birth†	Mono	2.13	3.16	3.80	3.00	3.08
	Biling	1.74^{ns}	3.16^{ns}	3.81^{ns}	3.35^{ns}	2.84^{ns}
Grade	Mono	11.51	8.00	10.36	10.94	9.74
	Biling	10.13^{ns}	9.00^{ns}	9.84^{ns}	10.65^{ns}	9.41^{ns}
Income	Mono	2.23	1.71	2.38	2.29	2.19
	Biling	1.86^{ns}	1.77^{ns}	1.78*	1.95^{ns}	1.86**
Test	Mono	19.98	17.66	39.23	26.41	27.42
	Biling	14.21*	14.35^{ns}	37.84^{ns}	23.00	21.12**
Sample Size	Mono	51	38	61	17	173
	Biling	53	43	32	20	155

* = t-test significant at .05 level
** = t-test significant at .01 level
ns = t-test not significant at .05 level
† - coded by decade

Interestingly, these differences were not mirrored for all groups individually except for test scores. Thus, although there is a vague pattern evident which could be interpreted, it certainly is not definitive. Generally, respondents interviewed by monolinguals appear to be somewhat more competent in English and somewhat more affluent than those interviewed by bilinguals. As with the child data, it is impossible to tell whether the results are due to a response bias or a sampling bias. In the former case the hypothesis is that individuals answer differently to monolinguals than to bilinguals, while in the latter case one would assume that the difference lies in the people for whom interviews were and were not completed; that is, monolinguals may complete a higher proportion of interviews with respondents who have a better command of English and who have higher incomes, while bilinguals may complete a higher proportion of interviews with respondents knowing little English and with small incomes.

To the extent that this is a viable explanation, it is worth elaborating on its implications for the SIE. The principal reason for "incompleted" interviews in the field test was that the individuals to be interviewed could not be found. Sometimes the address was non-existent or the family was not known at the address. In other cases, the individual to be interviewed was temporarily out of the area or had moved without leaving a forwarding address. To a large extent, an interviewer's rate of interview completions was a function of his or her ability to "track down" the respondent. How monolingual and bilingual interviewers might have differed at this task in the field test is moot presently, and may be irrelevant to the SIE in any case since the SIE interviewers will be assigned to addresses rather than specific people. This should minimize the non-response rate due to inability to locate the appropriate respondents.

3. Performance of the Monolingual and Bilingual Data in a Discriminant Function -

As one final analysis, the child data collected by the matched monolingual and bilingual interviewers were placed in the discriminant function derived using list as criterion and recommended in Chapter VII. The resulting LESA and non-LESA categorizations were then matched against list categorizations for those children. The results are given in Table 4. Subsequent tests showed that none of the pairs of percents differed significantly from one another. Thus, it can be concluded that there is little evidence of systematic effect of interviewer type on LESA - non-LESA classification.

Table 4: Performance of list discriminant function when used on data collected by monolingual vs. bilingual interviewers: Child data, list as criterion.

	Cubans		Chicanos		Navajos		Chinese		Overall	
	Mono	Bil	Mono	Bil	Mono	Bil	Mono	Bil	Mono	Bil
% classified the same by MELP and list	79	67	83	84	64	71	80	73	78	74
% classified LESA by List	66	67	48	47	70	61	44	54	57	56
% Classified LESA by MELP	72	76	58	56	70	54	48	42	63	59
% Bias	+9	+15	+21	+19	0	-12	+9	-21	+11	+5

4. Summary

While this substudy did not show large differences in data collected by monolinguals and bilinguals, its design had two weaknesses relative to its implications for the SIE:

1. Monolinguals were generally more experienced at interviewing than bilinguals. Apparently, RTI did not match the five pairs in each site for experience. Therefore, we may be comparing data collected by experienced monolinguals with those collected by inexperienced bilinguals.
2. The list sampling procedure resulted in only 60-65% response rate. Thus, the results of this study confound two factors: (a) differential skills in locating respondents, a skill not relevant to the SIE. (b) Differences in answers to MELP questions given by respondents to bilingual vs. monolingual interviewers.

In view of these problems, the results of the monolingual-bilingual comparisons are not definitive in any sense.

Bibliography

Bowen, J. D. Measuring language dominance in bilinguals. Working Papers in Teaching English as a Second Language, 1974, 8, 13-32.

Burt, M. H. Dulay, H., and Hernandez-Chavez. Bilingual Syntax Measure. New York: Harcourt, Brace and Jovanovich, 1974.

Capco, C. S., and Tucker, G. R. Word association data and the assessment of bilingual education programs. TESOL Quarterly, 1971, 5, 335-342.

Chomsky, N. Aspects of the theory of Syntax. Cambridge, Massachusetts: M.I.T. Press, 1965.

Clarke, J. Theoretical and technical considerations in oral proficiency testing in R. Jones and B. Spolsky (eds.) Testing Language Proficiency, Arlington, Virginia: Center for Applied Linguistics, 1975.

Cochran, W. G., and Hopkins, C. E., Some classification problems with multivariate qualitative data. Biometrics, 1961, 17, 10-32.

Cohen, Andrew. Sociolinguistic assessment of speaking skills in a bilingual education program. In L. Palmer and B. Spolsky (eds.) Papers on Language Testing, 1967-1974. Washington, D.C.: 1975.

Committee on Irish Language Attitudes Research. Report submitted to the Minister for the Gaeltacht, October 1975.

Cramer, Harold. Mathematical Methods of Statistics. Princeton, New Jersey: Princeton University Press, 1946.

Dailey, John. Language Facility Test. Alexandria, Vriginia: The Allington Corporation, 1968.

Darnell, D. K. Clozentropy: A Procedure for testing English language proficiency of foreign students. Speech Monographs, 1970, 37, 36-46.

Dunkin, M., and Biddle, B. The Study of Teaching. New York: Holt, Rhinehart and Winston, 1974.

Educational Testing Service. Test of English as a Foreign Language. Princeton, New Jersey: Educational Testing Service, 1974.

Foreign Service Institute. Absolute Language Proficiency Ratings (Circular). Washington, D.C.: Foreign Service Institute, 1963.

Fishman, J., Cooper, R., and Ma, R. Bilingualism in the Barrio. Bloomington Indiana: Indiana University Press, 1971.

Gradman, H., and Spolsky, B. Reduced redundancy testing: A progress report. In R. Jones and B. Spolsky, Testing Language Proficiency, Arlington, Virginia: Center for Applied Linguistics, 1975.

Guttman, Louis. A basis for scaling qualitative data. American Sociological Review, 9, 1944, 139-150.

Harrison, W., Prator, C., and Tucker, G.R. English-Language Policy Survey of Jordan. Arlington, Virginia: Center for Applied Linguistics, 1975.

Herbert, C. The Basic Inventory of Natural Language. San Diego, California: Chess and Associates Inc., 1975.

Hymes, D. Models of the Interaction of language and social setting. Journal of Social Issues, 1967, 23, 8-28.

Ilyin, Donna. Ilyin Oral Interview. Rowley, Massachusetts: Newbury House, 1972.

Jones, R. L., and Spolsky, B. Testing Language Proficiency. Arlington, Virginia: Center for Applied Linguistics, 1975.

Kelly, L. G. (Ed.) Description and Measurement of Bilingualism: an International Seminar, Toronto, Ontario: University of Toronto Press, 1969.

Kirk, S., McCarthy, J., and Kirk, W. The Illinois Test of Psycholinguistic Abilities. Revised edition. Urbana University of Illinois Press, 1968.

Labov, W. The logic of nonstandard English in F. Williams (Ed.), Language and Poverty. Chicago: Markham, 1970. 153-189.

Levenston, E. A. Aspects of Testing the Oral Proficiency of Adult Immigrants to Canada. In L. Palmer and B. Spolsky (Eds.), Papers on Language Testing, 1967-1974. Washington, D.C.: TESOL, 1975.

Lieberson, S. How Can We Describe and Measure the Behavior of Bilingual Groups? In L. G. Kelly (Ed.), Description and Measurement of Bilingualism: An International Seminar, Toronto, Onterio: University of Toronto Press, 1969.

Mackey, W. F. How Can Bilingualism be Described and Measured? In L. G. Kelly (Ed.), Description and Measurement of Bilingualsim: An International Seminar. Toronto, Ontario: University of Toronto Press, 1969.

Macnamara, J. How can one measure the extent of a person's bilingual proficiency? In L. G. Kelly (Ed.), Description and Measurement of Bilingualism: An International Seminar. Toronto, Ontario: University of Toronto Press, 1969.

Matluck, J., and Matluck, B. MAT-SEA-CAL Oral Proficiency Measure. Seattle, 1975.

Mehrabian, Albert. Nonverbal Communication. Chicago: Aldine, 1972.

Menyuk, P. A preliminary evaluation of grammatical capacity in children. Journal of Verbal Learning and Verbal Behavior, 1963, 2, 429-439.

Naiman, N. The use of elicited imitation in second language research. Working Papers in Bilingualism, 1974, 2, 1-37.

Natalicio, D., and Williams, F. Repetition as an oral assessment technique. Austin, Texas: Center for Communication Research, 1970. (ERIC # 051680).

Nie, N. et al. 1975. Statistical Package for the Social Sciences. New York: McGraw-Hill Inc.

Oller, J. Dictation as a device for testing foreign language proficiency. English Language Teaching, 1971, 25, 254-259.

_____. Scoring methods and difficulty levels for cloze tests of proficiency in English as a second language. Modern Language Journal, 1972, 56, 151-158.

Oller, J. and Streiff, V. Dictation: A test of grammar based expectancies. In R. Jones and B. Spolsky (Eds.), Testing Language Proficiency. Arlington, Virginia: Center for Applied Linguistics, 1975.

Rosenshine, B., and Furst, N. Research on teacher performance criteria. In B. Smith (Ed.), Research in Teacher Education: A Symposium. Englewood Cliffs, New Jersey: Prentice-Hall, 1971.

Scott, M. S. Error analysis: A study of Arab students' written and oral production in English: Unpublished M.A. thesis. American University of Beirut, 1973.

Simon, A., and Boyer, E. (Eds.), Mirrors for behavior: An anthology of classroom observation instruments. Philadelphia: Research for Better Schools Inc. 1967.

Spolsky, B. Reduced Redundancy as a language testing tool. In G. Perren and J. Trim (Eds.), Applications of Linguistics: Selected Papers of the Second International Congress of Applied Linguistics, Cambridge, 1969. London: Cambridge University Press, 1971.

Stemmler, A. The LCT, Language Cognition Test (Research Edition) - A Test for Educationally Disadvantaged School Beginners. In L. Palmer and B. Spolsky (Eds.), Papers on Language Testing. 1967-1974. Washington, D.C.: TESOL, 1975.

Taylor, W. L. Cloze Procedure: A new tool for measuring readability. Journalism Quarterly, 1953, 30, 414-438.

Upshur et al. Michigan Test of Language Proficiency. Ann Arbor, Michigan: English Language Institute, 1964.

Upshur, John. Objective Evaluation of Oral Proficiency in the ESOL classroom. TESOL Quarterly, 1971, 5, 47-60.

Wilds, Claudia. The Oral Interview Test. In R. Jones and B. Spolsky (Eds.), Testing Language Proficiency, Arlington, Virginia: Center for Applied Linguistics, 1975.

Wolfram, W., and Fasold, R. The Study of Social Dialects in American English. Englewood Cliffs, New Jersey: Prentice-Hall. 1974.

APPENDIX 1

Letter to Center for Applied Linguistics from National Center for Education Statistics requesting a proposal for research and development activities leading to a Measure of English Language Proficiency.

DEPARTMENT OF HEALTH, EDUCATION, AND WELFARE
OFFICE OF EDUCATION
WASHINGTON, D.C. 20202

Dr. Rudolph C. Troike, Director
Center for Applied Linguistics
1611 North Kent Street
Arlington, Virginia 22209

Dear Dr. Troike:

On behalf of the National Center for Education Statistics, this office would be pleased to receive from the Center for Applied Linguistics a technical proposal to develop a validated measure of the Census for use in its survey of children counted for purposes of Title I, ESEA. The Census Bureau Title I survey is mandated by P. L. 93-380, Sec. 822(A), and the survey of limited English-speaking ability among persons from non-English language backgrounds is mandated in Sec. 731(c)(1)(A) of the same Public Law. Design specifications for the measure(s) to be developed may be found in the attachment.

The due date for all final products for use by the Bureau of the Census is October 3, 1975. The final report to NCES incorporating all technical materials, full documentation, evidence of reliability and validity of the measures developed and tested, minutes of several advisory group meetings representing the linguistic, "research," and ethnic communities, and all other products to be agreed upon mutually may be submitted at a later date, but not later than March 31, 1976. Submit each product first in (at least one) draft and allow the NCES up to five working days for review. Naturally, given the "tight" deadlines, you may expect much quicker response; NCES will have available at all times a project monitor and an associate to expedite its review.

The technical proposal should contain the following:

1. Introduction. This should contain a concise discussion demonstrating your understanding of the problem of developing a measure of limited English-speaking ability acceptable to the Bureau of the Census.

2. Work plan. In this section of the proposal there are specific descriptions of how you plan to design, establish and implement the development program on a task-by-task basis. The proposal should clearly state how you intend to proceed to identify and develop measurement alternatives, to design the "test sites", to arrange for development on-site, to compare and evaluate measure alternatives, and to document the recommended measure fully. The proposal must be exceedingly clear on how the Center for Applied Linguistics intends to work with NCES to relate jointly to the Bureau of the Census to produce the specific products to be delivered for use by the Bureau of the Census. The proposal should show how the CAL would establish system evaluation criteria and parameters, obtain and use information required for evaluation of measures and arrive at recommendations. The technical proposal should demonstrate that the work plan would produce a measure with the desired properties and in a form (items, ratings, training materials, etc.) manifestly acceptable to the Bureau of the Census and the NCES. The plan should be comprehensive, going well beyond the information contained in the statement of design specifications. A Pert chart or other comparable plan for outlining the essential steps to be conducted within the scope of this procurement, their approximate duration and products to be delivered should be included in this part of the proposal.

3. Personnel.

   A. Vitae of all key professional project personnel. Specific qualifications related to the proposed project should be noted. Examples of previous work relevant to this project by key personnel should be indicated (with identification of sponsor and monitor) and should be available upon request.

   B. Names, qualifications, and responsibilities of consultants and subcontractors. (CAL is encouraged to utilize as consultants minority professionals and as subcontractors minority-owed firms with special capabilities relevant to work in bilingual education. Also be certain to include in the staffing at least one mathematical statistician with experience designing studies or experiments for survey-related work.)

4. Management plan. The proposal shall include a detailed statement describing plans to organize, staff and manage the project. It is estimated that the equivalent of approximately four or five professional man-years of effort will be required, exclusive of the costs of producing videotapes and renting playback equipment in sufficient quantities (if indicated) and costs of convening advisory groups for the work to be carried out.

BEST COPY AVAILABLE

The plan should include a schedule by phase and tasks. An organizational chart should also be submitted indicating the relationship of the project team to the organization. The technical proposal should provide a staffing plan by phase and task with a table or chart showing each key individual or category of support staff to be employed on the project, descriptions of the tasks which each individual will perform, the periods of time during which each task will be performed, the number of person-days estimated for each individual for each task, and total for estimated person-days by individual and by task. (This same staffing plan should be included also in the separate cost proposal.)

The separate cost proposal should repeat the staffing plan from the technical proposal, in identical format, and show the dollar cost for each individual for each assignment. Daily or hourly rates of pay for each person must be quoted. An itemized detailed budget is required, including documentation of the overhead costs. Costs for subcontracts included in the budget should be separately itemized. In addition, the costs and time estimated to be incurred for ADP personnel such as programming and computer analysts, should be identified by task. While the cost of the computer facility at DHEW will be borne by the Government, CAL is requested to estimate the costs of the usage of the DHEW computer in terms of dollars or CPU minutes by phase. If the proposal suggests the use of an outside computer for the processing of the data collected in the field, the estimated cost should be specified.

Because the time to develop the measure under this proposed procurement is rapidly running out, I would appreciate receiving your proposal at the earliest possible opportunity, but no later than close of business, Thursday, May 15. At that time we will want 12 copies of the technical proposal and 3 copies of the cost proposal. Send them to me at Room 1077, 400 Maryland Avenue, S. W., Washington, D. C. 20202.

This letter is not to be construed as a contract award nor will your response to this letter obligate the Government to make an award to you on the basis of your proposal.

If I may be of further assistance to you during the preparation of your proposal, please feel free to call me at 245-8630.

Sincerely,

Jacob J. Maimone, Chief
Research, Development and Statistics Branch
Grant and Procurement Management Division

Attachment

APPENDIX 2

Design specifications for MELP by Dr. Burton R. Fisher

Final Report

DESIGN SPECIFICATIONS FOR A MEASURE OF LIMITED ENGLISH-SPEAKING ABILITY IN A NATIONAL SURVEY

Prepared for
National Center for Educational Statistics
Education Division
Department of Health, Education, and Welfare

by

Burton R. Fisher

717 Knickerbocker Street
Madison, Wisconsin 53711

April 1975

The studies for this report were conducted pursuant to Contract No. P00-75-0057 with the Office of Education, U. S. Department of Health, Education, and Welfare. Contractors undertaking such projects under Government sponsorship are encouraged to express freely their professional judgment in the conduct of the project. Points of view or opinions stated do not, therefore, necessarily represent official Office of Education or National Center for Educational Statistics position or policy.

DESIGN SPECIFICATIONS FOR A MEASURE
OF LIMITED ENGLISH-SPEAKING ABILITY

Some General Boundaries

1. Various DHEW policy documents, and pages 148-149 of the Conference Report on HR 69, make it clear that we are concerned with measurement of English language ability, and not with language dominance or with proficiency in any language other than English.

A second principal and fixed requirement, for a host of good reasons, is that the national survey of "limited English-speaking ability" (LESA) mandated in Sec. 731 (c) (1) (A) of Title VII ESEA be carried out by the Bureau of the Census for NCES. It will be "piggybacked" on Census's large-scale national survey of the economic status of families, mandated in Sec. 822 (a) of PL 93-380.

One may not find the latter measurement context requirement optimal, but the function of the R & D work for which design features are specified here is to find solutions within the constraints set forth, and not to raise problems.

Some More Specific Boundaries

2. Census people say that if measurement of LESA is to be carried out in the Census survey, at least four constraints must be observed.

a. "Testing" in any overt form, identifiable by respondents as such, is definitely excluded; this applies especially to "paper-and-pencil" tests. This places a limit on the kinds of response-eliciting stimuli which can be used to get at LESA.

b. Also categorically excluded is electronic recording of what the respondent says, for later analysis and coding. This places a limit on the kinds of responses to be recorded and the locus of assessment of these responses.

c. A third explicit constraint: LESA measurement procedures must not break rapport during the interview, must fit "naturally" into the context and content of a CPS-like interview (face-to-face or via telephone), and must be within the capacity of its usual CPS and CPS-like interviewers. (On the whole, the latter are women 35 - 40 years of age, with a high school education.) The procedures must not disrupt them.

d. The strong preference of the Census staff is for as simple a measure as is feasible, with a small series of direct questions, answerable by the usual respondent for the household about all of the other memberd of the household. (In about 60% of CPS interviews, this is the mother.) That is, the preference is for enumeration of the household members, without sampling within the household to select the actual respondents.

This is a strong Census preference, not an absolute requirement. Whether this preference can be gratified, given the need for an adequate measure of LESA (a key NCES requirement), is an empirical questions to be answered in the course of R & D work.

Some Design Specifications

3. The LESA measurement (survey) was mandated by Congress in PL 93-380. That it be done as adequately as possible, that it be accepted as a legitimate measure, while bending the above Census requirements as little as possible is a central R & D task. In view of the attitudes of the serveral Congressional, public and educational constituencies involved, the adequacy of the measure of LESA is in a broad sense a "political" matter -- apart from the requirements of the professional standards at NCES and Census. The specifications for R & D work proposed below, and the several considerations set forth which enter into the specifications, have the above constraints and professional standards in mind.

BEST COPY AVAILABLE

a. Labels: This is not to be a test of "limited English-speaking ability." It is to be a "measure of English language proficiency." It is not to be a measure of English language competence or aptitutde for learning English; it is to be a measure of English language performance and mastery, as they appear in a defined measurement situation. Let us call it MELP, for present purposes. It can have alternative forms.

b. Sec. 703 (a) of Title VII ESEA defines: "The term 'limited English-speaking ability', when used with reference to an individual, means..."individuals who "have difficulty speaking and understanding instruction in the English language" because "they were not born in the United States or whose native language is a language other than English" or because they "they come from an environment where a language other than English is dominant. Further, "The term 'native language', when used with reference to an individual of limited English-speaking ability, means the language normall used by such individuals, or in the case of a child, the language normally used by the parents of the child."

Other references in Pl 93-380 (to preschool education; to auxiliary and supplementary programs for parents of LESA pupils; to elementary and secondary education; to bilingual education under the Adult, Vocational and Higher Education Acts), and the language of Sec. 731 (c) mandating this survey make it clear that the "individuals" referred to above may be of any age. However, individuals aged 5 - 17 seem to be of special interest.

Furthermore, the definitions of "program of bilingual education" for Title VII ESEA and the several educational Acts cited above indicate that the Congress holds that these programs of instruction are appropriate and necessary because the LESA (of those whose native language is not English or who come from foreign-language dominant environments) is a barrier to the effective progress of their education and training. It is primarily for these persons of LESA that federally-supported programs of bilingual education are intended.

c. From the words and sentences of PL 93-380, the following interpretations and inferences may be drawn:

(1) In the survey, MELP is to be obtained only for persons of the defined demographic and language community characteristics. (For the moment, we put aside consideration of whether or not comparison data from those in groups defined by other characteristics ought to be obtained in the survey and/or MELP R & D work.)

This would involve a series of "screening" questions addressed to the usual respondent for the household in Census surveys. Furthermore, it may turn out that is desirable practically to have these questions administered by unselected Census interviewers for "screening" purposes, with a more complex version of MELP (see below) later administered by a more highly trained interviewer. As will be seen below, when the validity-standardization study is discussed, some of these questions and a few additional simple questions may also be useful and more easily administered surrogates for MELP in its more elaborated form.

The formulation of these "screening" questions is not a simple matter at all, and there is considerable controversy as to the nature of language questions in Census work. (See Lieberson, 1966, and others.) Under these circumstances, it would be highly desirable that this set of questions be prepared by the R & D contractor in close association with Census people. Experience with the Bilingual Supplement to the July 1975 CPS should be helpful in this work.

(2) MELP is to be an individual measure, except for very young children (where it is to be derived from the "screening" items). It is an empirical question, for R & D work, as to whether the usual single Census informant about the members of the household can validly and reliably provide MELP data of sufficiently discriminatory power about the household's individuals.

If the answer is "no", there will probably be need for sampling and

BEST COPY AVAILABLE

interviewing respondents within the household selected in the Census sample. Particularly when the thus-selected individual for interview is a child, or other than the available and usual respondent for a household, there will be further R & D work called for. This would involve instruction and training for interviewers in maintaining rapport and minimizing the potential awkwardness of this novel situation. From survey experience, it may be said that the interviewer's typical fear of such awkwardness is likely to be very much greater than the actual difficulty.

(3) The phrase "...speaking and understanding instruction in the English language..." is interpreted to mean oral production (encoding in speech) and aural comprehension (decoding others' speech) in English. In the several education statutes, when reading and writing have been in mind the sophisticated statute drafters have seen fit to specify them directly; such specification is absent here. The consequences of this interpretation would be that:

(a) "Paper-and-pencil" writing tests and presentation of printed materials for reading can be kept out of the interview, accomodating to a Census constraint.

(b) If direct questions about how well an individual speaks and how well an individual understands English, put to that individual or to someone else about him, yield unsatisfactory MELP data, there is an alternative approach. The individual's speaking and understanding behaviors may be observed during the course of the interview itself, in response to questions which at least overtly do not appear to attempt to elicit either a range of language behaviors or an assessment of language behavior by the respondent.

(c) Given the Census veto of electronic recording of the respondent's language behavior (for later extra-interview analysis by experts), the interviewer may be trained to record and assess/rate behaviors he has been cued to watch for on forms developed by the R & D work on MELP. This is not an unusual procedure in good psychological and social research and in assessment work in organizations. People without previous expertise and special qualifications have been successfully trained to make reliable and accurate reports and assessments of behaviors during group interactions and individual performances, in field and in laboratory situations.

(4) Whatever form of MELP is used, we are faced with a choice. It must either be universally applicable to individuals of all age/educational levels -- perhaps with accompanying age/educational level norms for interpretation of individual results; or it must involve different versions of the measure , perhaps with different measurement procedures, for individuals of different age/educational levels or for individuals differentially accessible for face-to-face interview.

This is both a theoretical and empirical matter, to be ascertained during MELP R & D work. The simplest form of MELP which has adequate measurement characteristics is the ultimately preferred form.

(5) We continue with our inferences and interpretations, drawn from the language of PL 93-380 and applied to our design specification purposes. Thus, the word "instruction" in the definition of LESA, and the strong implication in the definitions of bilingual education programs -- that they are intended to promote effective education and training for those now disadvantaged by LESA in schooling carried out in English -- lead to the following conclusion. MELP must in some direct or indirect way not only be content (performance) - referenced, but it must also be criterion-referenced. Whatever the form of the MELP, we must know that we have a valid measure -- one whose "scores" accurately discriminate among individuals who have been identified as actually not making or making effective progress in their education and training by virtue of LESA. If the version of MELP developed for use is content-referenced and samples English language behavior adequately, the validity of the measure and the establishment of cut-points can be reinforced by a consensus of expert opinion.

BEST COPY AVAILABLE

(a) On validity; MELP is to measure what it is intended to measure -- the characteristics and relative proficiency of "speaking and understanding instruction in the English language," which make a difference or could make a difference in the individual's progress in a course of education or training. How "limited" ESA is, for present purposes, is to be referenced against the language performance of individuals whose ESAs are seen by the schools as barriers of varying strength to effective learning, when instruction is in English.

(b) This applies to individuals in (preschool?), elementary, secondary, postsecondary, adult and vocational education programs. MELP validity studies ideally should be carried out in all of these contexts.

(c) It should be recognized that different educational agencies (SEAs, LEAs), schools and programs use different measures and criteria (of different worth in terms of scientific standards) both of ESA and of effective educational progress. The procedures for identifying individuals for whom LESA is in varying degrees a barrier to utilizing effectively instruction in English will thus also differ. The R & D contractor may be able to make some choices among these educational sites, as to where MELP developmental and validation studies should be carried out. (The modes of stratification for a purposive sample of sites in which to carry out such studies is left for later consideration by the R & D contractor.)

(d) For both practical and theoretical reasons, we are not likely to arrive at a "true" (essentially metaphysical) definition and measure of characteristics and degrees of ESA which universally ought to facilitate or inhibit educational attainment. We can obtain administrative identifications, in the schools as they are and by the identification methods they currently use, of individuals inhibited from normal educational attainment by LESA. This is a ubiquitous problem in research on exceptionalities, and the approach suggested here echoes experiences derived from that research.

(e) Were we to have a sufficiently large and differentiated sample of educational sites, from sub-sample data we could establish regional, institutional characteristics and (within the former groupings) age/grade level reference points for degrees of ELP related to probability or ease/difficulty of effective educational progress. It is questionable whether, for the purposes of the present national survey, such differentiated standards are desirable -- or even possible to obtain in an R & D study of reasonable dimensions. From the MELP data obtained at the total sample of educational sites in the validity study, and from their review by an expert group, national "cut-points" for LESA and MELP could be established -- for different age groups, at least.

Estimating the numbers of LESA persons of various characteristics by SEA or LEA or other boundaries is an issue and a procedure separable from the question of separate regional and other standards.

4. The sites of MELP validation are simultaneously proposed as the sites of MELP construction, particularly for what we shall call MELP's elaborated form. The intention would be to develop an"instrument"to measure ELP suited to the Census survey procedures while reasonably modifying them. This must always be kept in mind.

a. Specialists in applied linguistics have knowledge of the components and dimensions of phonology (accents, sounds, some dialect features), of lexicon, of syntax and of utterances to be used to characterize oral production and aural comprehension. (Parenthetically: Bilingual interviewers or non-verbal behavioral response indicators may be necessary, where an individual comprehends but does not speak English.) Applied linguists are aware of certain central "diagnostic" linguistic features of adequate and inadequate English language usage and comprehension. If they do not already know which of these linguistic features are most highly correlated with other features of English language usage, they can determine this empirically in R & D work at the educational sites. (The purpose of this is to shorten the list of language behaviors to be observed for entering into an assessment of ELP made by trained interviewers. The aim is practical -- while maintaining

BEST COPY AVAILABLE

a list of critical items long enough for MELP reliability.)

b. The next steps would be to prepare:

(1) tentative "ordinary question" unobtrusive standard stimuli, likely to elicit the speech production features to be observed in oral production responses;

(2) when these linguistic features are used in the standard stimuli, they bring forth overt behaviors or speech in English (assuming interviewers are not bilingual) indicating aural comprehension.

c. Tentative observed language behavior recording and rating/assessment recording forms (cueing the interviewer as to what kinds of behaviors to observe) would be prepared. These forms are likely to contain some combination of sets of qualitative categories, ordinally ordered categories, and continuous (but actually ordinal) "scales."

d. Tentative eliciting stimuli and response reporting/rating techniques would be applied "blind" by the R & D team at the validation sites, to individuals administratively designated by the schools as functioning with varying degrees of LESA (including zero) which interferes with education and training to varying extents. Various selected age/grade-level and different dominant non-English language individuals would be given these measures, the findings being treated separately -- at least in this try-out stage.

The key matter to be ascertained is how well which elements of the tentative MELP discriminate among the categories of LESA-identified individuals. A summary MELP "score" for oral production, and another one for aural comprehension would be derived and validated as above. It may even be possible to develop "scores" for finer features or sets of features of the individual's language behavior.

(1) The individuals discussed above should be those defined in PL 93-380 as possessing the specified demographic/language community "screening" characteristics. One check on elaborated MELP would be to apply it to individuals who lack these characteristics (e.g., born in U. S. English monolinguals whose parents speak only English), in the same sites.

(2) These procedures, improved in successive trials, would in later stages be employed with observer/raters who are Census-type interviewers trained by the R & D team as it develops its training operations. The stopping-point for R & D work would be signaled when a valid MELP relatively adequately meeting psychometric standards for intra- and inter-interviewer reliability and discriminatory power has been developed. How finely MELP should discriminate the quality of English language oral and aural mastery is left open; as a practical matter, it will probably be critical that the finest and most reliable discriminations be made in the central range of MELP "scores", where instruction-inhibiting LSA transits to instruction-barring LSA.

(3) The elaborated MELP version thus developed (and alternative versions) would receive their final validation in education and training sites of the various kinds <u>other</u> than those utilized for MELP development. The reasons for this are obvious.

(4) Finally, the developed versions of MELP must be pretested in the field, in a realistic CPS-like context in cooperation with the Bureau of the Census -- and revised as is necessary. If MELP has alternative versions, this is the opportunity to gain information as to which version is the "best" or "least bad" under the simulated conditions of the national survey.

e. On training of Census-type interviewers for using MELP: It will be necessary to prepare R & D interviewer training materials suitable for later relatively standardized training of Census field staff (regular or specially recruited) during a comparatively short training period carried out at dispersed locations. (Note: The CPS interviewer field staff meets for training at several central locations each month. Since the national survey will extend over several months, there should be opportunity for interviewer retraining and training reinforcement.)

248 BEST COPY AVAILABLE

For this purpose, videotaping of behaviors, made during the validation study -- or perhaps by professional actors following scripts -- and videotape cassette reproduction is proposed. What would be required, among other things, are:

(1) Videotaped examples of a variety of language behaviors, clearly displaying the linguistic features to be observed in oral production and the indicators of aural comprehension -- whether the latter be non-verbal action, or non-English speech addressed to a bilingual interviewer, or a response in English. Accompanying each sight-sound example would be a didactic discussion of what features have been displayed, how they are to be categorized and assessed, what they must not be confused with, etc. The examples would show individuals of various ages and various high-frequency English deficits and accomplishments -- whose primary language is not English.

(2) Videotaped exercises -- relatively discrete segments of oral production and aural comprehension behaviors, followed by full MELP field interviews, would be shown. The trainees would be asked to make their categorizations or ratings or other assessments (including a "global" assessment of ELP) on the standard forms. The trainer would then give the "correct" answers and how they were arrived at -- all still on videotape. A trainer would be available in person to answer questions and to receive "feedback" from the trainees. Both MELP and the specifications for its field administration, as well as the training program, can profit from such "feedback" -- if experience is to be our guide.

(3) It is possible that an entire training session presented on videotape for the trainees to observe could have unique training value, in addition to the more active processes described above. We are familiar with "sing-alongs"; why not a "measure-along"?

(4) The selection and preparation of material, pretesting and other activities in connection with the training program constitute an R & D study of itself. Again, advice and cooperation from Bureau of the Census personnel seem called for.

(5) Accompanying the preparation of videotaped training materials is the preparation and pretesting of clear written instructions for MELP use, which the interviewers can refer to in the field. (A toll-free number for the interviewer to call for advice, if she meets with difficulties in using MELP, would not be amiss.) In a sense, the interviewer's task then is to compare and assess actual respondent behaviors against reference standards and examples learned in training and described in the written instructions.

5. The development of one or more versions of the elaborated MELP described above is intended to produce the linguistically and psychometrically "best" performance measure of English language proficiency tied in with educational performance -- one whose quality and relevance will be legitimated by professional and public opinion.

On the other hand, it is reasonable to ask: Are there other measures which can be developed, psychometrically relatively respectable, correlating relatively highly with both elaborated MELP and the validity criterion, which possess certain advantages over elaborated MELP? Among these advantages might be: considerably shorter and less complex interviewer training required, no need for bilingual interviewers, less interview time consumed, less potential interview disruption, simpler data processing, and in general less trouble for the Bureau of the Census and its survey operations.

That is, can we develop measures simpler than elaborated MELP which are technically "good enough"? Can we trade off some technical quality and quantity of information for much greater operational ease, and still have a sufficiently reliable, valid and useful MELP? There is not complete assurance that a technically adequate elaborated MELP acceptable to the Bureau of the Census can be developed; there is a good chance of success in these respects. The issues raised above are really empirical questions, to be answered in R & D work. In any case, elaborated MELP must be there if the answers to these empirical questions are in the negative.

What more-or-less cumulative set of simpler measurement approaches might be explored?

a. Extending the range of "screening" questions to include ascertaining the possible use of English in various domains of language use (home, peers, work, etc.) and for various communication functions (e.g., radio and TV listening to English-language stations, re aural comprehension). These questions would probably be put to the usual CPS respondent for the entire household and about its individual members, and could include items on specific kinds of difficulties individuals might have in oral production and aural comprehension.

b. Ratings of household members, individually, on how well they speak and how well they understand English speech, made by the single respondent for the entire household.

c. Items equivalent to a. and b. above, where the respondent reports about and rates himself or herself; the individuals have been selected by within-household sampling. (There is even some point in a 100% sample of the household "cluster", where the household was itself selected in a probability sample -- though this would pose some practical problems.)

d. CPS-type interviewers, with short and simple training, categorize/rate the respondent on how well the person speaks and how well he understands English -- and possibly whether his ELP is sufficient to effectively utilize an age-appropriate educational or training opportunity. In the normal course of an interview, the interviewer has had an opportunity to observe the language behavior of the respondent, and is supplied with appropriately cued reporting forms. She can ask direct questions.

e. In R & D work, it may be feasible to obtain a variety of demographic and language characteristics of the respondent who rates and categorizes persons within her household, and similar data about the interviewer. From these data, and the corresponding simple and elaborated MELP data, an appropriate "correction factor" might be applied to the results of the simpler MELP version to decently estimate what that measure's value would be on elaborated MELP.

f. Some combination of a. to e. above.

6. A rather different approach would be to ascertain simply-obtained predictors of the individual's elaborated MELP status and/or predictors of administratively identified LESA status at the validation sites. Some of the predictors might be the "screening" question responses of the informant for the household's members; others might be of the kind suggested in 5. a. to d. above. Still others might be the usual Census demographic data on household members and data on the household as a unit. A multiple regression equation, whose regressors are obtainable in a household interview of the CPS variety, yielding reliable and accurate estimates of the elaborated MELP or LESA status dependent variables, would be the goal. This could be one of the distinctive tasks of R & D work.

7. The MELP produced in R & D work should as far as is possible meet the technical and other criteria set forth in the 1974 revision of Standards for Educational and Psychological Tests. It would be beyond the function of this design statement to rehearse these standards.

8. The R & D team is envisioned as being composed of specialists experienced in applied linguistics, in several aspects of psychometrics, in educational practices concerning LESA students, and in survey work as conducted by the Bureau of the Census. (A Census professional as liaison person with the R & D team is a minimum requirement.) As far as possible, the staff should include members of the major language communities.

9. Close association with the Bureau of the Census is emphasized for a series of reasons which affect the appropriate form for MELP.

a. Census people indicate that they have greater freedom of action with respect to interviews at households included in supplementary samples, compared with the constraints on interviews at regular CPS panel households. This greater flexibility pertains to interview content and procedures, and to the possibility of within-household respondent selection. There will be contingencies in the sampling plan for the Sec. 822 (a) PL 93-380 national survey and accommodations of aspects of that plan

to the needs of the LESA national survey. These contingencies will have implications for the MELP form used.

b. Another contingency is the language competence of Census interviewers. For the very large-scale surveys under discussion here, NCES has been given to understand that additional interviewers will be hired. It is apparently not an entirely closed question as to whether bilinguals can or will be specified for hire. Census could also be asked to ascertain how many of its current interviewers are bilingual, in what languages other than English, and where located geographically. The bilingual interviewer permits a simpler form of the measure of aural comprehension of English (while posing some problems in the accuracy of assessment of oral production in English. Further, should Census specify that a certain proportion of the interviews be conducted *via* telephone, bilingual interviewers become even more essential. For MELP activity, face-to-face interviews are greatly to be desired.

10. PL 93-380 provides an excellent roster of the many kinds of public and professional constituencies interested in the national survey of LESA, and its implications for bilingual education planning and programs. The communities of linguists and psychometricians are also involved. All of these groups, in some advisory capacity to NCES (and by extension to the R & D contractor) can provide the kinds of legitimations helpful to acceptance of both MELP and the national survey.

APPENDIX 3

Narrative of Principal Activities of MELP Project: June 1975 - June 1976

Appendix 3

Project Narrative

The following notes summarize the principal activities of the project during each of its phases:

1. Instrument Development and Refinement (Chapter III):

June 2 - 8: Stolz and Troike went to San Francisco to meet Ms. Minerva Mendoza-Friedman to recruit her as the project's San Francisco coordinator. They also met with Harold Yee, president of Asian, Inc., who advised them on renting office space and making contacts in the various ethnic communities. In addition, they met with Ms. Teresa Chen and Prof. Susan Ervin-Tripp, both of University of California - Berkeley to initiate recruiting efforts for research assistants and junior research assistants.

Strick and Jones reviewed possible assessment instruments in Arlington, and recruiting of LGRs and the planning of the first LGR meeting continued.

June 9 - 15: The San Francisco office was established and nine research assistants began work on June 12. Strick took charge of developing discrete point tests, and contacts with the local ethnic communities were established to begin recruitment of households in which to try out various instruments. Initial versions of instruments were produced. On June 9, Stolz and Strick consulted with Dr. Charles Herbert of Chess and Assoc., author of the Basic Inventory of Natural Language, about the possibility of using the BINL as a criterion instrument.

Initial meetings of the LGRs were held in Arlington June 10 - 19. The schedule was as follows:

June 10 - 11	Spanish Speakers
June 12 - 13	Native Americans
June 14 - 15	Chinese

June 16 - 17 Asian/Pacific Group

June 18 - 19 Europeans

The agenda and proceedings are attached to this report. Generally, each group was oriented to the project and the SIE. They reviewed some tentative instruments for assessing English proficiency. They made suggestions about specific instruments and/or items that they thought would or would not work in their groups. They also recommended various interviewing techniques. Representatives of NCES and RTI were present.

On June 13, Roger Shuy, Director of Domestic Programs at CAL, briefed the Federal Interagency Language Roundtable on the project.

June 16 - 22: Leslie Silverman, Project Monitor for NCES, and Michael Weeks, Director of interviewer training for RTI, joined the San Francisco staff and began an extended discussion of the field test design which lasted essentially the entire week. Silverman and Stolz met with Harold Yee who suggested that the validation of instruments be carried out within a "known groups" design using pre-identified LESA and non-LESA samples. This notion was carried back to the design meetings and formed the basis of most of the discussion. On June 18, Dr. John Upshur of the University of Michigan joined the group as a specialist in testing language proficiency. He had been a consultant during the writing of the proposal. On June 19, Troike and Burton Fisher arrived and joined the discussion.

During this time the research assistants continued to test preliminary versions of discrete point tests in the three ethnic communities. Also, a number of junior research assistants were recruited.

In Arlington, Dr. Jeanne Freeman of CAL began developing a behavior observation system for use by monitors in observing interviewer-respondent interactions during the field test.

June 23 - 29: Twelve junior research assistants joined the staff, and interviewing using trial instruments began in earnest. The staff divided itself into groups with each group concentrating on the development of a particular instrument. On June 27 Stolz and Troike held a briefing for Federal Education Community representatives in Arlington.

June 30 - July 6: Upshur returned to take temporary charge of the San Francisco activities while Stolz was not on site. Silverman also returned and began working with a group of research assistants on drafts of the MELP questions. On June 30 and July 1, a group of language assessment specialists composed of Ms. Clandia Wilds of Washington, D. C. (creator of the FSI Oral Interview), Protase Woodford of Educational Testing Service, Edward D'Avila of Bilingual Children's Television, Dr. Evelyn Hatch of U.C.L.A., and Sidney Sako of Defense Language Institute reviewed the progress of instrument development to data and made the suggestion that additional effort be placed on the development of a direct interviewer-rating system for use in the field test. Stolz returned to San Francisco on July 2 and began the development of the Direct Observation Rating Procedure (DORP).

Freeman came to San Francisco to begin testing of the monitoring system.

July 7 - 12: The San Francisco activities centered on:

1. Development of the DORP
2. Analyzing data collected using trial versions of various instruments, with subsequent elimination of poor items or entire tests.
3. Preparing "final" versions of the MELP questions, discrete point tests, and DORP for review by OMB and the LGRs at their second meeting.
4. Training staff on the monitoring system using videotapes of interviews recorded earlier in the week.

July 13 - 18: On July 13 - 14 the second LGR meeting was held in San Francisco. LGRs were briefed on the progress of the project and then given copies of the instruments. Members of the project staff role-played interviews with LGRs to

familiarize them with the materials and procedures. Feedback, criticisms, etc. were solicited from each LGR. Representatives of NCES, Census, and RTI were in attendance.

The San Francisco operation was then shut down and all field-test materials underwent final reworking by Strick and RTI's staff to prepare for training RTI supervisory personnel on July 18.

2. Field Testing the Instruments (Chapter IV):

July 22 - 24: Interviewer training in El Paso and Miami

July 25 - August 16: Data collection in El Paso and Miami

July 29 - 31: Interviewer training in Arizona and San Francisco

August 1 - 23: Data collection in Arizona and San Francisco

3. Data Analysis:

September 3 - 4: LGR Meeting #3, Arlington, Va.

September 22 - 24: A conference of experts was held in Arlington to choose the questions to be recommended as the MELP questions. (Chapter V)

October 2: A memorandum was delivered to NCES recommending the set of questions to be used in the SIE as the MELP. The memo did not deal with the question of how to map responses to the questions on to LESA and non-LESA categories.

October 3 - March 30, 1976: Statistical Analyses were done focused on the production of scoring keys for converting answers to MELP questions into LESA and non-LESA categorizations (Chapters VI and VII).

March 30: Contract extended to June 15, 1976 at no additional cost to the government.

April 5 - 6: Conference of specialists to consider recommendations for additional activities to recalibrate and/or revalidate the MELP, using data collected in the SIE. (Chapter XI)

Participants included:

Dr. John B. Carroll- University of North Carolina
Harold Yee- Asian, Inc., San Francisco
Rosa Inclan- Dade County Public Schools
Burton Fisher- University of Wisconsin
Dr. Daniel Horvitz- R.T.I.
Dr. Tyler Hartwell- R.T.I.
Leslie Silverman- NCES
Dr. Dorothy Waggoner- NCES
John Conway- NCES
Dr. Lepa Tomic- O.C.R.
Roy Rodrigues- O.C.R.
Carter Holling- N.I.E.
Michael Rand- Bureau of the Census
Marvin Thompson- Bureau of the Census

A report of that meeting is appended to this report.

April 1 - 30: Analysis of bilingual-monolingual interviewer effects and first-hand versus proxy responses to MELP questions. (Chapters IX and X)

April 22: Presentation of preliminary MELP project results to American Educational Research Association (this dissemination activity was not supported by Government funds).

May 1 - June 15: Preparation of final report.

APPENDIX 4

Tests considered but not used as criterion measures.

Appendix 4.

Accounts of Work with Discrete-Point Criterion Measures which were not Included in the Final Test.

Tests considered but not field tested. 1. The Bilingual Syntax Measure (Burt, Dulay and Hernandez-Chavez, 1974) was considered for a test of production. It was developed to test a child's oral proficiency in English and is an example of a discrete point, indirect test. The child is shown several cartoon-like pictures and asked a series of questions about them. The questions are constructed to elicit specific grammatical structures by the child. There are 25 items on the test, and it takes 10 to 15 minutes to administer. The scoring is very simple: one simply counts the number of grammatically correct answers.

Although the test has many good features, it was not further considered for two reasons. First, it was not applicable to children over 9 years. Second, the test would have been relatively expensive to use. (The retail price of the kits would have been over $4000.)

2. Dailey Facility Test. This test (Dailey, 1968) was also considered for a test of oral production for children. It is not a discrete point test, but rather an integrative direct test. The child is shown a series of pictures (representing different domains - school, home, playground) and asked to tell a story based on each picture. There is no time limit. The stories are recorded. Later a rating of 0 to 9 is given to the story. The following is a description of these ratings.

9 ....A well-organized story with imagination and creativity. Need not be original. May use well-known fictional or historical characters.

8 ....A complete story, but not a well organized one.

7 ....An interpretation of some elements of implied action or intentions, as deduced from or suggested by the picture but not a complete story.

6 ....A detailed description of what is happening, but nothing about past or future action or intentions. At level 6 all or nearly all of the elements of the picture will be covered, in contrast to level 5 where only some selected elements will be covered.

5 ....A partial description consisting of two or more sentences with some description of movement or action as seen in the picture.

4 ....Two or more sentences describing persons or objects but no verb of action or indication of interaction between a person and an object.

3 ....A complete sentence that makes sense.

2 ....Compound responses, two or more words at a time, a single word describing action, or more than one single-noun response.

1 ....One single-noun response.

0 ....No response -- garbled speech, or only pointing at picture.

The test was dropped from further consideration for two reasons. First, the pictures were unsuitable: many were culturally biased; others were too sophisticated for children. Second, the rating system was too ambiguous. It was felt that it could not be used reliably without much interviewer training and further development of the scoring system.

3. The Basic Inventory of Natural Language (BINL). This test was considered for a test of oral production. It was developed by Charles Herbert (1975) to measure a child's oral language dominance and proficiency. Children are trained to tell stories (based on a set of visual materials) to their peers. The stories are recorded and later transcribed. A set of 10 utterances are then selected for analysis. They are scored for fluency (the average number of words per utterance) and syntactic complexity (different weights are given to utterances with full sentences, partial sentences, phrases, and clauses). The test thus falls into the discrete point direct category.

Although the elicitation technique used in this procedure was very appealing, the test was not used for pilot-testing for two reasons. First, it could not be assured that there always would be a second child in the household to whom the target could narrate a story. Second, it was felt that the scoring procedure lacked clear face validity.

Measures which were piloted and then dropped. As explained above, these were tests that were fielded in San Francisco, and then completely eliminated from the battery. There were two such tests.

1. Word Naming. This test was developed by Fishman, Cooper, and Ma (1971) to measure bilingual proficiency and is an integrative indirect measure. Basically the respondent was asked to name as many different words as possible which were found in a particular domain. For example, he was given 1 minute to name in English objects found in the home. Other domains were school, neighborhood, and work. This procedure was also repeated in Spanish. Fishman found high correlations between the number of words given and the most frequently used language in the home. There was also a high negative correlation between the number of English words and a Spanish literacy factor.

The test was adapted in the following ways for our purposes. It was used as a test of oral production and was only given in English. It was administered to both children and adults. Each respondent was asked to name objects in 3 domains. Adults were asked to name objects found at home, in the neighborhood and at work. They were given one minute for each domain. Similarly, children were asked to name objects found at home, in the neighborhood and at school. The score for each respondent was the total number of different and contextually appropriate object names (see Appendix 11 for instructions and questions).

The test was dropped from the battery because of the difficulty of controlling the testing situation. That is, it was found that the subject would often look around the room where he was tested and name the objects present. Thus scores were a function of the "business" of the room in which the subject was tested. Because not enough time was available to modify the technique, or to standardize the situation, the test was eliminated from the battery.

2. ETS Listening Comprehension Test. This unpublished test was originally developed by ETS for the Puerto Rican Ministery of Education to test students' level of achievement of certain curriculum materials. As will be seen CAL adapted this test to measure English receptive and productive ability in children and adults.

The test had four levels:

Level 1 was given to children in grades 1-3.
Level 2 was given to children in grades 4-6.
Level 3 was given to children in grades 7-9.
Level 4 was given to students in grades 10 and above.

Levels 1,2, and 3, had two sections. In Part 1 the subject was shown 4 pictures. The examiner said a sentence (e.g. There is a spoon on the table) and asked the subject to point to the best picture. In Part 2, the subject was shown 4 pictures and read a short passage. The examiner then asked him a question about the passage. The subject was required to point to the most appropriate picture (e.g. A boy broke Jane's bicycle. Her father fixed it, and she helped him by handing him the tools he needed. What was broken?)

Level 4 only had one section which corresponded to Part I described above. Each test had the following number of items.

	Part 1	Part 2	Total
Level 1	50	10	60
Level 2	45	10	55
Level 3	50	20	70
Level 4	70	none	70

The total score was simply the number of correctly identified pictures. The test was clearly a discrete point indirect one.

In the San Francisco pilot work CAL made the following major modification. For each item described above, not only was the respondent required to point to the appropriate picture but he also had to say the answer. In the case of Part I, this meant repeating the sentence said by the examiner. In Part II, the respondent was required to verbally answer the question. Thus each item was scored for information and grammar. In Part 1 a number of crucial structures were identified in each sentence. If these were correctly repeated the subject would receive a point. The number of structures varied from sentence to sentence, some had one (the boy hit the ball), some had more (That boy wants to play baseball). A point was given for each correctly repeated target structure. In case the response was a totally grammatical alternate, the respondent was given only one point in addition to the possible point for identification. In part 2, a point for correct grammar was given only if the information in the sentence was correct as well. The answers did not have to be complete sentences.

The test was given to both children and adults. Forms were selected by age rather than grade, thus if a 20 year-old subject only had a grade 5 level education he was given Level 4 rather than Level 2.

As the pilot work progressed, items were eliminated from the tests when they appeared to be culturally inappropriate or did not discriminate good from poor speakers. (See Appendix 11 for various forms and developments of the test).

Eventually the test was entirely eliminated. The pre-emptive reason was that CAL had to receive permission from the Puerto Rican government in order to use it. This process would have been too lengthy and complicated. There were also other problems with the test: each level was too long; the scoring of the production part

was troublesome. That is, a respondent might correctly repeat the target structures, but make mistakes in other parts of the sentence and still receive a perfect score.

APPENDIX 5

Window Rock Analyses

The Window Rock Data

We have already mentioned that the school lists were constructed in Window Rock based solely on the students' scores on the comprehension section of the Gates-McGinitie Reading Test. Moreover, the assignment to lists was done purely on the basis of grade level; i.e. if a student's comprehension score was below his grade level, he was placed on the "low" or LESA list, otherwise he was placed on the "high" or non-LESA list.

Examining the data from Window Rock, it became immediately clear that the list information was not appropriate for our purposes. Consider Tables 1a and 1b below. Table 1a shows the relationship between test score, in terms of total points correct, and grade level, while Table 1b shows the relationship between list membership and grade level.

Table 1a: Window Rock Children by grade and test score.
Test: total points

Grade	0-30	31-50	51-67	Total
K-3	7	21	18	46
4-6	1	22	66	89
7-8	0	3	33	36

Table 1b: Window Rock Children by grade and school list
List

Grade	LESA - below grade	Non-LESA at or above grade	Total
K-3	9	37	46
4-6	54	35	89
7-8	27	9	36

If test score is taken as the measure of English proficiency, Table 1a supports the hypothesis that, by and large, the older children know more English

than the younger ones. This general pattern was replicated in all other sites, regardless of whether test score or list was used as a measure of English proficiency. However, Table 1b would indicate just the opposite: that the older the child is, the less he knows of English. This is a truly abberrant pattern given all of our data and what is known about second language acquisition. The problem characterized in Table 1b, therefore, seems to be peculiar to reading and not to English proficiency. That is, it appears that the Window Rock children rapidly fall behind nationally normed grade levels in reading comprehension as they grow older. However, the conclusion that this is due to a decrease in their English proficiency appears not to be tenable.

On the basis of these data, we decided not to use the Window Rock list information in deriving our scoring keys. Thus, when list was used as a criterion variable, only the data from Ganado were utilized. Of course, when test scores were the criterion measure, the data from all Navajo children were combined into a single sample.

APPENDIX 6

Regression Analysis - Children

Regression Analysis - Children

At an early stage in the analysis of the field test data, a series of multiple regression analyses were performed for both descriptive and analytical reasons. Later, however, it became clear that discriminant analysis was more to the point of this project and that the regression analyses added nothing to it. Thus, these analyses did not result in a scoring key. The basic results of the multiple regression analyses will be briefly presented below for those who are accustomed to thinking about multi-variate prediction problems such as the present one in regression terms.

Table 1 presents the regression analyses within each group and for all groups pooled using the ten MELP variables as predictors and FCTR as the criterion. Coefficients denoted as B are unstandardized while those denoted as β are standardized.

Table 1 : Children's Data: Results of Regression Analysis with FCTR as Criterion.*

	Cuban		Chicano		Chinese		Other Asian		Navajo		Overall	
Sample size	317		364		146		133		260		1220	
R	67		73		66		60		61		67	
R^2	45		53		43		37		37		45	
Variable	B	β	B	β	B	β	B	β	B	β	B	β
WHEN	-02	-01	26	13	02	03	-02	-05	-54	-06	02	02
SPEAK	131	14	18	20	02	03	-06	-15	19	28	16	19
UNDERSTAND	22	23	10	10	14	22	21	49	-05	-07	11	13
H LANG	-23	-05	12	10	-11	-12	06	12	14	18	05	05
SIB	06	04	09	08	06	08	04	14	-03	-04	08	08
FRND	15	13	29	24	17	21	-09	-12	15	19	21	20
YEARS	18	28	18	13	05	09	04	20	-05	-12	12	20
BIRTH	-03	-07	-08	-07	-12	-20	-01	-07	01	04	-04	-08
GRADE	02	04	-01	-01	03	08	02	10	18	45	01	02
PARENT	-02	-02	05	07	02	04	01	06	03	05	03	05
CONSTANT	14		178		700		51		-100		08	

* All coefficients given to two decimal places, decimal points omitted.

271

APPENDIX 7

Regression Analysis - Adults

Regression Analysis - Adults

The adult data were subjected to multiple regression analysis using the 11 MELP variables as predictors and FCTR (not dichotomized) as criterion. Table 1 gives the regression analysis as performed within each ethnic group and across all groups.

Table 1: Adult Data: Results of Regression Analysis with FCTR as criterion*

	Cuban		Chicano		Chinese		Other Asian		Navajo		Overall	
Sample Size	272		202		111		116		214		915	
R	69		65		82		73		72		81	
R^2	47		42		67		53		53		65	
Variable	B	β	B	β	B	β	B	β	B	β	B	β
WHEN	18	11	-18	-16	10	07	14	16	-17	-03	10	08
SPEAK	22	28	15	20	35	46	10	14	07	11	18	25
UNDERSTAND	14	21	18	26	-09	-12	05	07	05	07	10	13
KID	-04	-03	03	03	-01	-01	-06	-08	11	14	02	02
FRIEND	-00	-00	-09	-07	13	10	04	07	04	05	04	03
HLANG	-29	-09	-13	-11	-03	-02	02	03	05	06	-03	-03
YEARS	01	02	09	27	06	31	01	12	03	15	04	21
NEWS	-13	-13	-22	-21	01	01	-25	-22	-31	-27	-18	-14
BIRTH	05	09	-01	-01	10	12	14	28	04	07	07	10
GRADE	04	23	01	04	04	17	02	15	03	17	03	15
INCOME	01	01	12	13	03	03	12	21	-07	-10	04	04
CONSTANT	-1.58		-0.68		-2.28		-1.04		0.24		-1.59	

* All coefficients given to two decimal places, decimal points omitted.

APPENDIX 8

Staff Utilization and Technical Consultants

STAFF UTILIZATION STATEMENT

Name	Position	Qualifications
Walter S. Stolz	Project Director	Ph.D. in mass communication, University of Wisconsin, Madison, Wisconsin. Formerly director of a project on child language acquisition - administrative experience as chairman of the psychology department at Earlham College - extensive teaching experience on statistics, research design, test design, psycholinguistics, and linguistics.
Rudolph C. Troike	Senior Project Advisor	Ph.D. in linguistics, University of Texas, Austin, Texas. Formerly director of the Texas Dialect Survey - administrative experience as chief administrator and policy manager of the Center for Applied Linguistics - has directed research in linguistics and anthropology.
Roger W. Shuy	Senior Project Advisor	Ph.D. in English linguistics, Western Reserve University. Formerly director of several field studies, including urban language surveys in Detroit and Washington, D.C. - administrative experience as Associate Director for Domestic Programs at CAL and Director of the Sociolinguistic Program at Georgetown University.
Margaret Bruck	Statistical Analyst	Ph.D. in experimental psychology, McGill University, Montreal, Quebec, Canada. Research in bilingualism, language development tests, and measurement experience in experimental design and statistical analysis.
Edward Fuentes	Statistician	M.S. in psychology and counseling, California State University, Los Angeles, California. Ph.D. ducational psychology, Stanford University, Stanford, California. Experience in statistical analysis, language test construction, and identification of bilingual children language: Spanish

Name	Title	Qualifications
Robert Pearl	Statistical Advisor	M.A. economics and statistics, American University, Washington, D.C. Formerly Chief of Economic Statistics Branch of the Population Division, Bureau of the Census - extensive experience in data collection in household surveys in several Federal Agencies.
Jack Upshur	Test Construction Specialist	Ph.D. in psychology, University of Michigan, Ann Arbor, Michigan. Administers external language testing programs, English Language Institute, University of Michigan, Ann Arbor, Michigan. Extensive experience in testing and publications.
Margaret Griffin	Staffing Consultant	Ph.D. in sociolinguistics, Georgetown University, Washington, D.C. Has extensive nationwide contacts with graduate students appropriate for field staff. Research Staff Linguist at CAL.
Minerva Mendoza-Friedman	San Francisco Coordinator	M.A. in educational psychology, School psychologist with San Francisco Unified School District. Contacts with Latin American community Language: Spanish. Panamanian background.
Jeanne Freeman	Junior Staff - Research Assistant	M.A. American Civilization, University of Texas, Austin, Texas. Ph.D. in curriculum and instruction, University of Texas, Austin, Texas. Research in developing coding systems for evaluations; teaching experience.
Ted Jones	Junior Staff - Research Assistant	B.A. in psychology, Earlham College, Richmond, Indiana. Teaching assistant in statistics and methodology.
Gregory Strick	Junior Staff - Research Assistant	M.A. Education/TEFL, American University, Beirut, Lebanon. Ph.D. candidate in sociolinguistics, Georgetown University, Washington, D.C. Teacher of English - research on English-Arabic code-switching and error analysis. Language: Arabic. Field experience in U.S., Canada and Lebanon.

Name	Title	Qualifications
Leann Parker	Junior Staff-Project Coordinator	M.A.-University of Texas, Austin, Texas. Has a background in linguistics and education; has had experience working with community and educational groups as well as experience in corrdinating projects and meetings at CAL. Language: Spanish.
Gilbert N. Garcia	Junior Staff-Coordinator of Language Group	Ph.D. student in applied linguistics. Is a sociolinguist with experience in U.S. and foreign field research; has also had experience in working with multi-lingual/multi-cultural community groups. Language: Spanish. Mexican-American background.
William Leap	Junior Staff-Advisor to Indian LCR Group	Ph.D. in anthropology, Southern Methodist University, Dallas, Texas. Staff Research Associate on Indian education, CAL. Extensive experience with various Indian education programs and contacts with various Indian groups across the U.S.
Amador Bustos	Research Assistant/Monitor	B.A. in philosophy and Chicano Studies, University of California, Berkeley, California. M.A. candidate, Chicano Studies, University of California, Berkeley, California. Research on Mexican Americans and language attitudes. Language: Spanish.
Anna Lai	Research Assistant/Monitor	M.A. in school psychology, San Francisco State University. Community work with recent Chinese immigrants and with mental health center. Language: Cantonese. Chinese background.
Alberto Rey	Research Assistant/Monitor	Ph.D. Spanish. Research in Spanish Linguistics - experience in translation - familiarity with Miami, Florida Cuban community. Language: Spanish. Cuban background.
Pedro Ruiz	Research Assistant/Monitor	B.A. in sociology, University of California, Berkeley, California. M.S.W. in Community Organization, University of California, Berkeley, California. D.D., Hastings College of the Law, San Francisco, California. Extensive experience in social work related organizations - work in Mexican-American community. Language: Spanish. Panamanian background.

Name	Title	Qualifications
Michael SamVargas	Research Assistant/ Monitor	M.A. in psychology, U.C.L.A., Los Angeles, California. Ph.D. candidate in social psychology, Harvard University, Cambridge, Massachusetts. Research on political and racial attitudes, Chicano psychology work in California Mexican American community. Languages: Spanish, Cantonese. Spanish and Chinese background.
John T. Webb	Research Assistant/ Monitor	Ph.D. in Spanish, U.C.L.A., Los Angeles, California. Teaching experience in ESL and Spanish. Translator and interpreter - research on U.S. Spanish. Languages: Spanish and other romance languages, German, Russian Mexican American brakground.
Benjamin Zambaler	Research Assistant/ Monitor	B.S. in Business Administration, Philippine School of Business Administration, Manila, Philippines. M.S.W. student, San Francisco State University, San Francisco, California. Volunteer counselor in Philippine community. Languages: Philippine languages.
Ophelia Balderrama	Junior Research Assistant/Monitor	B.A. student in sociology, San Francisco University, San Francisco, California. Program Coordinator for Mission Adult Center and other community centers. Worked with Spanish speaking adults and children in San Francisco. Language: Spanish. Mexican-American background.
Richard Chambers	Junior Research Assistant/Monitor	M.A. candidate in applied linguistics, University of California, Berkeley, California. Research and teaching experience in bilingual education and ESL. Language: Spanish.
Elizabeth C. Dunigan	Junior Research Assistant/Monitor	B.A. in English education, College of Notre Dame, Belmont, California. Elementary school teacher. Experience in art and photography (assisted with project videotaping).
Evangeline Kamitsuka	Junior Research Assistant/Monitor	B.A. student in psychology, Earlham College, Richmond, Indiana. Tutor - Counselor for Upward Bound Summer Program. Language: Japanese. Japanese background.

Name	Title	Qualifications
Carolyn Karelitz	Junior Research Assistant/Monitor	B.A. in sociology, New York University, New York. Graduate study in linguistics, University of California, San Diego, California. Experience in field work - research on Polynesian languages.
Cynthia Lindsey	Junior Research Assistant/Monitor	B.A. student in Spanish and elementary education, Earlham College, Richmond, Indiana. Community work in Spanish speaking community. Bilingual aide in an Albuquerque elementary school. Language: Spanish. Chinese, French, Cherokee, Blackfoot background.
Gloria Lozano	Junior Research Assistant/Monitor	B.A. student in modern languages, Earlham College, Richmond, Indiana. Spanish tutor. Languages: Spanish, French. Mexican-American background.
Roberta Mailman	Junior Research Assistant/Monitor	A.B. in linguistics, University of California, Berkeley, California. Certificate in TESL, U.C.L.A., Los Angeles, California. Teacher of ESL (elementary school). Research in linguistics (on non-native English speakers). Languages: Spanish, Hebrew.
Claire McKenzie	Junior Research Assistant/Monitor	B.A. in Spanish, U.C.L.A., Los Angeles, California. Certificate in TESL. Research in TESL and linguistics. Language: Spanish.
Annie Baguna Panlibuton	Junior Research Assistant/Monitor	B.A. student in psychology, San Francisco State University, San Francisco, California. Mental health worker in Philipino community in San Francisco, and other community work (theatre, youth). Languages: Tagalog, Visayan. Philippine background.
Margaret Robbins	Junior Research Assistant/Monitor	B.A. in German, Stanford University, Stanford, California. TESOL certificate from U.C.L.A., Los Angeles, California. Teaching experience in Germany.

Name	Title	Qualifications
William Sinclair	Junior Research Assistant/Monitor	B.A.-psychology, University of California, Berkeley, California. Chinese studies, Carnegie Chinese Institute, CSUSF, Ph.D. candidate in linguistics, University of California, Berkeley, California. Research linguist - African languages - experience with bilingual education program, San Francisco Unified School District. Languages: Spanish, Mandarin, Italian, and others.
Jennie Yee	Junior Research Assistant/Monitor	B.A.-psychology, San Francisco University, San Francisco, California. M.S. student in clinical psychology, San Francisco University, San Francisco, California. Community mental health work - work in Chinese community. Language: Chinese. Chinese background.
	Secretaries:	
Shirley Oravitz	Clerk-Typist	CAL
Marina Vargas	Clerk-Typist	San Francisco

TECHNICAL CONSULTANTS TO THE MELP PROJECT

Jere Brophy, (University of Texas, Austin, Texas; professor of curriculum and instruction; specialist in development of coding systems for verbal interaction)

John Carroll, (University of North Carolina; leading researcher in psycholinguistics; formerly with Educational Testing Service)

Andrew Cohen, (UCLA, Teaching English as a second language; bilingual education)

Robert Cooper, (School of Education, Hebrew University, Jeruselem, Israel, language testing)

Edward D'Avila, (Bilingual Children's Television, Berkeley; specialist in psycholinguistic research with Chicano children)

David DeCamp, (Associate Director, CAL; specialist in linguistics, English as a second language)

Burton Fisher, (University of Wisconsin, professor of Sociology, specialist in survey design and statistical analysis)

Joshua Fishman, (Yeshiva University, New York, N.Y.; specialist in social psychology and bilingual education)

John Francis, (Schoolmaster, Maret School, Washington, D. C.; language testing)

Gilbert N. Garcia, (Ph.D. student in applied linguistics and CAL staff - Spanish translator, Texas Mexican American background)

Evelyn Hatch, (UCLA; expertise in early childhood)

Charles Herbert, (Director, Chess, Inc.; Associated with University of California at Irvine; language test development specialist)

Ouillermo Hernandez, (Ph.D. candidate in Chicano studies, University of California, Berkeley, California; specialist in ethical analysis)

Rosa Inclan, (Director of Bilingual Education, Dade County Schools, Dade County, Florida; specialist in bilingual education)

Reynaldo Macías, (Ph.D. student in sociolinguistics at Georgetown University, Washington, D. C.; Spanish translator, California Mexican-American background)

Les Palmer, (American Language Institute, Georgetown University; test construction specialist in English as a second language; originally developed TOEFL -- Test of English as a Foreign Language -- while on CAL staff)

Robert Pearl, (Mid-Atlantic Research Institute, Bethesda, Maryland; survey research and design)

BEST COPY AVAILABLE

Alberto Rey, (Howard University, Washington, D. C.; professor of Spanish; Spanish translator, Cuban background)

Sidney Sako, (Defense Language Institute, San Antonio; Director for testing and evaluation for DLI)

Ivadnia Scott-Cora, (Howard University, Washington, D. C.; professor of Spanish, Spanish translator, Puerto Ricah background)

George Stanton, (Stanford University, graduate student; computer science)

G. Richard Tucker, (McGill University, Montreal, Quebec, Canada; professor of psychology; specialist in psycholinguistics and evaluation of bilingual programs)

John Upshur, (University of Michigan, Ann Arbor, Michigan; English Language Institute; specialist in English language testing)

Claudia Wilds, (D. C. Public Schools; originally refined methods used by Foreign Service Institute for oral language interview rating)

Protase Woodford, (Educational Testing Service; specialist in language testing, including Spanish and English as a foreign language)

Harold Yee, (President, Asian, Inc., San Francisco, California; specialist in statistical analysis)

Robert Young, (University of New Mexico, Albuquerque, New Mexico; specialist on Navajo language and culture)